




2021

## Scalable Approaches for Auditing the Completeness of Biomedical Ontologies

Fengbo Zheng

University of Kentucky, fzh229@gmail.com

Author ORCID Identifier:

 <https://orcid.org/0000-0001-5902-0186>

Digital Object Identifier: <https://doi.org/10.13023/etd.2021.128>

[Right click to open a feedback form in a new tab to let us know how this document benefits you.](#)

### Recommended Citation

Zheng, Fengbo, "Scalable Approaches for Auditing the Completeness of Biomedical Ontologies" (2021).  
*Theses and Dissertations--Computer Science*. 105.  
[https://uknowledge.uky.edu/cs\\_etds/105](https://uknowledge.uky.edu/cs_etds/105)

This Doctoral Dissertation is brought to you for free and open access by the Computer Science at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Computer Science by an authorized administrator of UKnowledge. For more information, please contact [UKnowledge@sv.uky.edu](mailto:UKnowledge@sv.uky.edu).

## **STUDENT AGREEMENT:**

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

## **REVIEW, APPROVAL AND ACCEPTANCE**

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Fengbo Zheng, Student

Dr. Mirosław Truszczyński, Major Professor

Dr. Zongming Fei, Director of Graduate Studies

SCALABLE APPROACHES FOR AUDITING THE COMPLETENESS OF  
BIOMEDICAL ONTOLOGIES

---

DISSERTATION

---

A dissertation submitted in partial fulfillment of the  
requirements for the degree of Doctor of Philosophy in the  
College of Engineering  
at the University of Kentucky

By

Fengbo Zheng

Lexington, Kentucky

Co-Directors: Dr. Licong Cui, Assistant Professor of Computer Science  
and Dr. Mirosław Truszczyński, Professor of Computer Science

Lexington, Kentucky

Copyright © Fengbo Zheng 2021  
<https://orcid.org/0000-0001-5902-0186>

## ABSTRACT OF DISSERTATION

### SCALABLE APPROACHES FOR AUDITING THE COMPLETENESS OF BIOMEDICAL ONTOLOGIES

An ontology provides a formalized representation of knowledge within a domain. In biomedicine, ontologies have been widely used in modern biomedical applications to enable semantic interoperability and facilitate data exchange. Given the important roles that biomedical ontologies play, quality issues such as incompleteness, if not addressed, can affect the quality of downstream ontology-driven applications. However, biomedical ontologies often have large sizes and complex structures. Thus, it is infeasible to uncover potential quality issues through manual effort. In this dissertation, we introduce automated and scalable approaches for auditing the completeness of biomedical ontologies. We mainly focus on two incompleteness issues – missing hierarchical relations and missing concepts. To identify missing hierarchical relations, we develop three approaches: a lexical-based approach, a hybrid approach utilizing both lexical features and logical definitions, and an approach based on concept name transformation. To identify missing concepts, a lexical-based Formal Concept Analysis (FCA) method is proposed for concept enrichment. We also predict proper concept names for the missing concepts using deep learning techniques. Manual review by domain experts is performed to evaluate these approaches. In addition, we leverage extrinsic knowledge (i.e., external ontologies) to help validate the detected incompleteness issues. The auditing approaches have been applied to a variety of biomedical ontologies, including the SNOMED CT, National Cancer Institute (NCI) Thesaurus and Gene Ontology.

In the first lexical-based approach to identify missing hierarchical relations, each concept is modeled with an enriched set of lexical features, leveraging words and noun phrases in the name of the concept itself and the concept’s ancestors. Given a pair of concepts that are not linked by a hierarchical relation, if the enriched lexical attributes of one concept is a superset of the other’s, a potentially missing hierarchical relation will be suggested. Applying this approach to the September 2017 release of SNOMED CT (US edition) suggested 38,615 potentially missing hierarchical relations. A domain expert reviewed a random sample of 100 potentially missing ones, and confirmed 90 are valid (a precision of 90%).

In the second work, a hybrid approach is proposed to detect missing hierarchical

relations in non-lattice subgraphs. For each concept, its lexical features are harmonized with role definitions to provide a more comprehensive semantic model. Then a two-step subsumption testing is performed to automatically suggest potentially missing hierarchical relations. This approach identified 55 potentially missing hierarchical relations in the 19.08d version of the NCI Thesaurus. 29 out of 55 were confirmed as valid by the curators from the NCI Enterprise Vocabulary Service (EVS) and have been incorporated in the newer versions of the NCI Thesaurus. 7 out of 55 further revealed incorrect existing hierarchical relations in the NCI Thesaurus.

In the third work, we introduce a transformation-based method that leverages the Unified Medical Language System (UMLS) knowledge to identify missing hierarchical relations in its source ontologies. Given a concept name, noun chunks within it are identified and replaced by their more general counterparts to generate new concept names that are supposed to be more general than the original one. Applying this method to the UMLS (2019AB release), a total of 39,359 potentially missing hierarchical relations were detected in 13 source ontologies. Domain experts evaluated a random sample of 200 potentially missing hierarchical relations identified in the SNOMED CT (US edition), and 100 in the Gene Ontology. 173 out of 200 and 63 out of 100 potentially missing hierarchical relations were confirmed by domain experts, indicating our method achieved a precision of 86.5% and 63% for the SNOMED CT and Gene Ontology, respectively.

In the work of concept enrichment, we introduce a lexical method based on FCA to identify potentially missing concepts. Lexical features (i.e., words appearing in the concept names) are considered as FCA attributes while generating formal context. Applying multistage intersection on FCA attributes results in newly formalized concepts along with bags of words that can be utilized to name the concepts. This method was applied to the *Disease or Disorder* sub-hierarchy in the 19.08d version of the NCI Thesaurus and identified 8,983 potentially missing concepts. We performed a preliminary evaluation and validated that 592 out of 8,983 potentially missing concepts were included in external ontologies in the UMLS.

After obtaining new concepts and their relevant bags of words, we further developed deep learning-based approaches to automatically predict concept names that comply with the naming convention of a specific ontology. We explored simple neural network, Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN) combined with LSTM. Our experiments showed that the LSTM-based approach achieved the best performance with an F1 score of 63.41% for predicting names for newly added concepts in the March 2018 release of SNOMED CT (US Edition) and an F1 score of 73.95% for naming missing concepts revealed by our previous work.

In the last part of this dissertation, extrinsic knowledge is leveraged to collect supporting evidence for the detected incompleteness issues. We present a work in which cross-ontology evaluation based on extrinsic knowledge from the UMLS is utilized to help validate potentially missing hierarchical relations, aiming at relieving the heavy workload of manual review.

KEYWORDS: Biomedical Ontology, Quality Assurance, Missing hierarchical relations, Missing concepts

FENGBO ZHENG  
\_\_\_\_\_  
Student's Signature

APRIL 30, 2021  
\_\_\_\_\_  
Date

SCALABLE APPROACHES FOR AUDITING THE COMPLETENESS OF  
BIOMEDICAL ONTOLOGIES

By  
Fengbo Zheng

LICONG CUI  
Co-Director of Dissertation

MIROSLAW TRUSZCZYNSKI  
Co-Director of Dissertation

ZONGMING FEI  
Director of Graduate Studies

APRIL 30, 2021  
Date

## ACKNOWLEDGEMENTS

I would like to thank the following people, without whom I would not have been able to complete my dissertation research, and without whom I would not have made it through my Ph.D. degree.

First and foremost, I would like to express my sincere gratitude to my advisor Dr. Licong Cui for her constant support and encouragement. It is Dr. Cui that lead me to this promising research direction. Dr. Cui's expertise was invaluable in formulating the research questions and methodology. Also, her insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

I would also like to thank the rest of my doctoral committee: Dr. Mirosław Truszczyński, Dr. Jinze Liu, and Dr. Sujin Kim. Especially, I really appreciate Dr. Truszczyński and Dr. Liu for handling all the official work at the University of Kentucky on behalf of Dr. Cui after Dr. Cui left UK.

I would like to thank the current and former Director of Graduate Studies Dr. Zongming Fei and Dr. Mirosław Truszczyński for their guidance throughout my Ph.D. study life.

I would also like to thank Dr. Debby Keen for sharing her teaching philosophy and preparing me to be a good instructor.

I would like to thank my colleagues Rashmie Abeysinghe, Jing Liu, Xufeng Qu, Yuanyuan Wu and my friends Liu Liu, Yu Zhao and Sifei Han for stimulating discussions as well as happy distractions to rest my mind outside of my research.

Finally, I would like to thank my parents for all their love and support. I know you are always there for me.



# Table of Contents

|  |            |
|--|------------|
| <b>Acknowledgements</b>  | <b>iii</b> |
| <b>List of Tables</b>  | <b>vii</b> |
| <b>List of Figures</b>   | <b>ix</b>  |
| <b>1 Introduction</b>  | <b>1</b>   |
| 1.1 Motivation . . . . .   | 1          |
| 1.2 Contributions . . . . .  | 2          |
| 1.3 Organization . . . . .   | 7          |
| <b>2 Background</b>  | <b>8</b>   |
| 2.1 Biomedical ontologies . . . . .  | 8          |
| 2.1.1 SNOMED CT . . . . .  | 9          |
| 2.1.1.1 Logical model . . . . .  | 9          |
| 2.1.1.2 Stated and inferred logical definitions . . . . .  | 11         |
| 2.1.2 National Cancer Institute Thesaurus . . . . .  | 12         |
| 2.1.3 Gene Ontology . . . . .  | 13         |
| 2.2 Techniques and extrinsic knowledge for supporting auditing . . . . .                                       | 14         |
| 2.2.1 Formal Concept Analysis . . . . .  | 14         |
| 2.2.2 Long Short-Term Memory (LSTM) . . . . .  | 16         |
| 2.2.3 Unified Medical Language System . . . . .  | 16         |
| 2.3 Quality assurance of biomedical ontologies . . . . .   | 17         |
| 2.3.1 Auditing concepts . . . . .  | 17         |
| 2.3.1.1 Auditing concept completeness . . . . .  | 18         |
| 2.3.1.2 Auditing concept modeling consistency . . . . .  | 19         |
| 2.3.2 Auditing hierarchical relations . . . . .  | 19         |
| 2.3.2.1 Structural-based approaches . . . . .  | 19         |
| 2.3.2.2 Lexical-based approaches . . . . .   | 20         |
| 2.3.2.3 Structural-lexical-based approaches . . . . .  | 21         |
| 2.3.2.4 Machine learning-based approaches . . . . .  | 21         |
| <b>3 A Lexical-based Approach for Exhaustive Detection of Missing Hierarchical IS-A Relations in SNOMED CT</b> | <b>23</b>  |
| 3.1 Methods . . . . .  | 23         |
| 3.1.1 Identifying stop words/phrases and antonym pairs . . . . .   | 23         |
| 3.1.2 Constructing lexical features for concepts . . . . .   | 24         |
| 3.1.2.1 Preprocessing FSNs of concepts . . . . .   | 24         |
| 3.1.2.2 Initializing lexical feature sets with noun phrases and words . . . . .                                | 25         |
| 3.1.2.3 Enriching lexical feature sets . . . . .   | 26         |
| 3.1.3 Identifying potentially missing hierarchical relations . . . . .   | 27         |

|          |   |           |
|----------|---|-----------|
| 3.2      | Results . . . . .   | 28        |
| 3.2.1    | Evaluation . . . . .  | 29        |
| 3.2.2    | Analysis of false positive cases . . . . .  | 29        |
| 3.3      | Discussion . . . . .  | 31        |
| 3.3.1    | Comparison with previous work . . . . .   | 32        |
| 3.4      | Conclusions . . . . .   | 35        |
| <b>4</b> | <b>Detecting Missing IS-A Relations in the NCI Thesaurus Using an Enhanced Hybrid Approach</b>  | <b>36</b> |
| 4.1      | Methods . . . . .   | 37        |
| 4.1.1    | Computing non-lattice subgraphs and generating candidate pairs  | 38        |
| 4.1.2    | Modeling concepts . . . . .   | 39        |
| 4.1.2.1  | Lexical features . . . . .  | 41        |
| 4.1.2.2  | Associative roles . . . . .   | 42        |
| 4.1.3    | Identifying potentially missing hierarchical relations . . . . .  | 43        |
| 4.2      | Results . . . . .   | 46        |
| 4.2.1    | Non-lattice subgraphs and suggested hierarchical relations . .  | 46        |
| 4.2.2    | Evaluation . . . . .  | 46        |
| 4.3      | Discussion . . . . .  | 47        |
| 4.3.1    | Analysis of false positives . . . . .   | 47        |
| 4.3.2    | Comparison with other approaches . . . . .  | 49        |
| 4.3.3    | Comparison with our previous work . . . . .   | 50        |
| 4.4      | Conclusions . . . . .   | 51        |
| <b>5</b> | <b>A Transformation-based Method for Auditing the IS-A Hierarchy of Biomedical Terminologies in the Unified Medical Language System</b> | <b>52</b> |
| 5.1      | Methods . . . . .   | 52        |
| 5.1.1    | Parsing concept names . . . . .   | 53        |
| 5.1.2    | Identifying replacement candidates . . . . .  | 54        |
| 5.1.3    | Concept name transformation . . . . .   | 55        |
| 5.1.4    | Identify missing hierarchical relations in source ontologies . . .  | 56        |
| 5.2      | Results . . . . .   | 57        |
| 5.2.1    | Identifying missing hierarchical relations . . . . .  | 57        |
| 5.2.2    | Evaluation . . . . .  | 58        |
| 5.2.3    | Analyses of false positive cases . . . . .  | 59        |
| 5.2.4    | Effect of restricting the hierarchical source for noun chunk replacement . . . . .  | 61        |
| 5.3      | Discussion . . . . .  | 62        |
| 5.3.1    | Distinction with related work . . . . .   | 62        |
| 5.3.2    | Exact versus normalized matching . . . . .  | 62        |
| 5.3.3    | Potential for concept enrichment . . . . .  | 63        |
| 5.3.4    | Applicability to a specific ontology . . . . .  | 64        |
| 5.4      | Conclusion . . . . .  | 64        |

|          |   |            |
|----------|---|------------|
| <b>6</b> | <b>A Lexical-based Formal Concept Analysis Method to Identify Missing Concepts in the NCI Thesaurus</b> | <b>65</b>  |
| 6.1      | Method . . . . .  | 65         |
| 6.1.1    | Constructing formal context . . . . .   | 65         |
| 6.1.2    | Identifying potentially missing concepts . . . . .  | 66         |
| 6.1.3    | Illustrative example . . . . .  | 67         |
| 6.2      | Results . . . . .   | 68         |
| 6.2.1    | Summary result . . . . .  | 68         |
| 6.2.2    | Preliminary evaluation . . . . .  | 68         |
| 6.3      | Discussion . . . . .  | 70         |
| 6.4      | Conclusion . . . . .  | 71         |
| <b>7</b> | <b>Exploring Deep Learning-based Approaches for Predicting Concept Names in SNOMED CT</b>               | <b>72</b>  |
| 7.1      | Method . . . . .  | 73         |
| 7.1.1    | Word embedding & data preprocessing . . . . .   | 73         |
| 7.1.2    | Neural networks for classifying word sequences . . . . .  | 74         |
| 7.1.3    | Predicting concept names given bags of words . . . . .  | 75         |
| 7.2      | Experiment & result . . . . .   | 76         |
| 7.2.1    | Experiment setup . . . . .  | 77         |
| 7.2.2    | Result for binary classification . . . . .  | 77         |
| 7.2.3    | Result for sequence prediction . . . . .  | 79         |
| 7.3      | Discussion . . . . .  | 81         |
| 7.3.1    | Potential Factors Affecting the Prediction Performance . . . . .  | 81         |
| 7.3.2    | Analysis of False Positives . . . . .   | 82         |
| 7.3.3    | Beyond Naming Purpose . . . . .   | 83         |
| 7.4      | Conclusion . . . . .  | 84         |
| <b>8</b> | <b>Preliminary Analysis of Cross-ontology Evaluation Based on Extrinsic Knowledge from UMLS</b>         | <b>85</b>  |
| 8.1      | Method . . . . .  | 86         |
| 8.2      | Result . . . . .  | 87         |
| <b>9</b> | <b>Discussion, Conclusions and Future Directions</b>  | <b>89</b>  |
| 9.1      | Discussion . . . . .  | 89         |
| 9.2      | Conclusions . . . . .   | 90         |
| 9.3      | Future directions . . . . .   | 94         |
| 9.3.1    | Repair Missing Hierarchical Relations . . . . .   | 94         |
| 9.3.2    | Improved Formal Concept Analysis . . . . .  | 94         |
| 9.3.3    | Deep learning approaches . . . . .  | 95         |
| 9.3.4    | Automatic validation method . . . . .   | 96         |
|          | <b>REFERENCES</b>   | <b>97</b>  |
|          | <b>Vita</b>   | <b>107</b> |

## List of Tables

|     |  |    |
|-----|--|----|
| 3.1 | The initial and enriched sets of lexical features of an example concept <i>c</i> : <i>371977004 – Primary malignant neoplasm of cecum (disorder)</i> . Noun phrases are underlined. . . . .  | 27 |
| 3.2 | Numbers of missing hierarchical relations detected in terms of the sub-hierarchies. . . . .  | 29 |
| 3.3 | Examples of missing hierarchical relations in the “ <i>Clinical finding</i> ” sub-hierarchy confirmed by the domain expert. . . . .  | 30 |
| 3.4 | Examples of false positives caused by the incorrect existing hierarchical relations. . . . .   | 31 |
| 4.1 | The role definitions of concept “ <i>Sarcoma</i> ” ( <i>C9118</i> ) in the NCI Thesaurus [3]. . . . .  | 37 |
| 4.2 | The number of potentially missing hierarchical relations identified for sub-hierarchies. . . . .   | 46 |
| 4.3 | Ten examples of valid missing hierarchical relations confirmed by EVS experts. . . . .   | 47 |
| 5.1 | An example of the transformation process for concept name “ <i>Acute dacryoadenitis of left eye.</i> ” . . . . .   | 55 |
| 5.2 | The number of potentially missing hierarchical relations detected in the UMLS source ontologies in English, as well as the ontology size and the number of existing hierarchical relations that can be identified for each ontology. . . . . | 58 |
| 5.3 | Examples of missing hierarchical relations confirmed by domain experts.  | 59 |
| 5.4 | Examples of false positives (or invalid missing hierarchical relations) and the existing hierarchical relations causing the false positives . . .  | 60 |
| 6.1 | The numbers of existing concepts, newly generated concepts, potentially missing concepts, and missing concepts validated via UMLS for each sub-hierarchy under <i>Disease or Disorder (C2991)</i> . . . . .                                  | 69 |
| 6.2 | Ten examples of validated missing concepts and their matched concepts in the UMLS ontologies. . . . .  | 70 |
| 7.1 | Result of binary classification for Experiment I. . . . .  | 78 |
| 7.2 | Result of binary classification for Experiment II. . . . .   | 79 |
| 7.3 | Result of LSTM-based sequence prediction in terms of the length of concept names. Training data is from September 2017 US Edition of SNOMED CT and test data is the newly added concepts in the March 2018 Edition. . . . .                  | 80 |
| 7.4 | Result of LSTM-based sequence prediction for names of missing concepts identified by Cui et al.’s method in [22]. . . . .  | 81 |
| 7.5 | Result of LSTM-based sequence prediction in terms of whether concept names contain duplicate words or not. . . . .   | 82 |

|     |  |    |
|-----|--|----|
| 8.1 | Ontologies and corresponding Path Contributions (PC) for the UMLS-based evaluation of detected subtype inconsistencies in Gene Ontology. | 88 |
| 8.2 | Ontologies and corresponding Path Contributions (PC) for the UMLS-based evaluation of detected subtype inconsistencies in NCI Thesaurus. | 88 |
| 8.3 | Ontologies and corresponding Path Contributions (PC) for the UMLS-based evaluation of detected subtype inconsistencies in SNOMED CT.     | 88 |

## List of Figures

|     |  |    |
|-----|--|----|
| 1.1 | An overview of the auditing approaches introduced in the dissertation.   | 4  |
| 2.1 | Diagram that shows the <b>stated</b> logical definition of concept <i>Upper back injury (disorder) (282765009)</i> [52] and notations of diagram elements [53]. . . . .  | 10 |
| 2.2 | Diagram that shows the <b>inferred</b> logical definition of concept <i>Upper back injury (disorder) (282765009)</i> [52]. . . . .   | 12 |
| 2.3 | Inferred defining relations and associations of concept <i>Benign Skin Neoplasm (C2896)</i> in the NCI Thesaurus [3]. . . . .  | 13 |
| 2.4 | Example formal context and the complete lattice formed by all formal concepts derived [62]. The formal context is of four geometry figures and four attributes. Formal concept $(\{1,2\},\{a\})$ is marked by dot in the formal context. . . . .   | 15 |
| 3.1 | The levels of concepts involved in a missing hierarchical relation: “ <i>Open injury of diaphragm (disorder)</i> ” IS-A “ <i>Open wound of thorax (disorder)</i> .”  | 32 |
| 3.2 | Distribution of potentially missing hierarchical relations detected in this work and previous work according to the level differences between subconcepts and superconcepts. . . . .   | 33 |
| 3.3 | A non-lattice subgraph identified in the previous work [26]. This non-lattice subgraph suggests a missing hierarchical relation between concepts 3 and 1: “ <i>Fracture subluxation of lunate</i> ” IS-A “ <i>Fracture dislocation of lunate</i> ”. . . . .  | 34 |
| 4.1 | An example of non-lattice subgraphs in the 19.08d version of NCI Thesaurus. Concepts are connected by hierarchical relations. The red dotted line shows a potentially missing hierarchical relation between concepts “ <i>Cutaneous Pseudolymphoma</i> ” and “ <i>Non-Neoplastic Skin Disorder</i> ” identified by our method. . . . . | 39 |
| 4.2 | Semantic models of concepts “ <i>Cutaneous Pseudolymphoma (C62776)</i> ” and “ <i>Non-Neoplastic Skin Disorder (C27555)</i> ” which are contained in the non-lattice subgraph shown in Figure 4.1. . . . .   | 40 |
| 5.1 | Dependency graph of the concept name “ <i>Primary basal cell carcinoma of left eyelid</i> .” . . . .   | 53 |
| 6.1 | Pseudocode of identifying potentially missing concepts by multistage intersection. . . . .   | 66 |

|     |   |    |
|-----|---|----|
| 6.2 | An example of FCA formal context generated by the concept <i>Breast Fibroepithelial Neoplasm (C40405)</i> in the NCI Thesaurus and its descendants in company with their lexical features. Word “Tumor” is normalized to “neoplasm” and word “Phyllodes” is normalized to “phyllode.” An FCA formal concept (marked by blue cells) with FCA attribute set {breast, neoplasm} is considered as a potentially missing concept among the given concepts. . . . . | 67 |
| 7.1 | Three neural network models used for the classification task. (a) is a simple neural network for binary classification; (b) is an LSTM neural network; (c) is the combination of convolutional neural network and LSTM. “None” means that the dimension is variable. . . . .  | 74 |
| 7.2 | Number of concept names in terms of the name length for all concepts in the March 2018 US Edition of SNOMED CT. . . . .   | 79 |
| 8.1 | <b>A:</b> An unlinked PMCP with diff 3 in SNOMED CT and its unlinked ITP derived; <b>B:</b> A linked PMCP with diff 3 in SNOMED CT and its linked ITP derived. This example reveals a potentially <b>missing hierarchical relation</b> in <b>A</b> , that is, <i>lesion of ligaments of shoulder region (disorder)</i> IS-A <i>soft tissue lesion of shoulder region (disorder)</i> . . . . .   | 86 |
| 8.2 | <b>A:</b> An unlinked PMCP with diff 3 in NCI Thesaurus and its unlinked ITP derived; <b>B:</b> A linked PMCP with diff 3 in NCI Thesaurus and its linked ITP derived. This example reveals a potentially <b>missing subtype relation</b> in <b>A</b> , that is, “ <i>connective tissues nevus</i> ” (CUI: <i>C8371</i> ) IS-A “ <i>connective tissue disorder</i> ” (CUI: <i>C26729</i> ). . . . .   | 87 |

## CHAPTER 1. Introduction

### 1.1 Motivation

An ontology (or terminology) provides formalized representation of knowledge within a domain, including a set of objects and the describable relationships among them. It provides a shared and common understanding of a domain that can be communicated between people and heterogeneous applications. In biomedicine, ontologies that ensure data consistency and interoperability, have played important roles in various biomedical research and applications, including biomedical data annotation, data integration and exchange, data analysis, information retrieval, natural language processing (NLP), and clinical decision support [1, 2]. For instance, the National Cancer Institute (NCI) Thesaurus [3], covering knowledge of cancers, genes and therapies has been widely used as a standard for biomedical coding, knowledge reference, and public reporting for many NCI and other systems [4–6]. SNOMED CT [7], the most comprehensive clinical healthcare terminology product in the world, facilitates the exchange of healthcare information among disparate healthcare providers and electronic health records (EHRs), leading to higher quality, consistency and safety in healthcare delivery [8, 9].

Biomedical ontologies are often incomplete and constantly evolving due to the growing knowledge in biomedicine, new requirements from emerging biomedical applications and the progressive nature of ontological engineering [10, 11]. Typically, ontology management involves the addition of new concepts along with their definitions and missing relations among concepts, as well as deprecation and deactivation of obsolete ones. For example, SNOMED CT is released regularly in every six months [12]. In the January 2019 release of SNOMED CT (International Edition), 11,903 new concepts were added and 3,035 concepts were inactivated; 20,294 changes were made to the stated concept definitions regarding relations between concepts. An-



other example is that NCI Thesaurus is updated every month with averaging roughly 700 new concepts in each release [13].

The lack of completeness, however, not only reduces the correctness and coverage of biomedical ontologies in modeling domain knowledge, but also affects the quality of downstream ontology-driven applications such as leading to valid conclusions being missed [11, 14]. For instance, in ontology-based search engines for patient cohort identification, queries are refined and expanded by moving up and down the hierarchy of concepts, thus incomplete ontology hierarchy will impact the quality of query results. As an example, value sets of SNOMED CT (consisting of subsets of SNOMED CT concepts) have been widely used for EHR decision support, quality reporting, and cohort selection. A value set can be defined as a list of concepts sharing some common features, e.g., all descendants of “*Carcinoma of larynx*”. However, “*Primary adenosquamous cell carcinoma of larynx*” is currently not listed as one of its descendants (i.e., missing hierarchical relation). Therefore, patients with “*Primary adenosquamous cell carcinoma of larynx*” will not be selected for the cohort of patients with “*Carcinoma of larynx*” which decreases the recall of the query result.

Due to the sheer size and complexity of modern biomedical ontologies, it becomes impractical to entirely rely on human effort to uncover quality issues, such as missing hierarchical relations and missing concepts [15]. Therefore, computational and automated quality assurance approaches are highly desired.

## 1.2 Contributions

Auditing the completeness of biomedical ontologies has been an active research area given its importance. Researchers proposed various methods to reveal quality issues in biomedical ontologies such as missing concepts [16–23] and missing hierarchical relations [22–32]. However, the ever-growing size and structural complexity make the quality assurance of biomedical ontologies a challenging task. As a result, existing

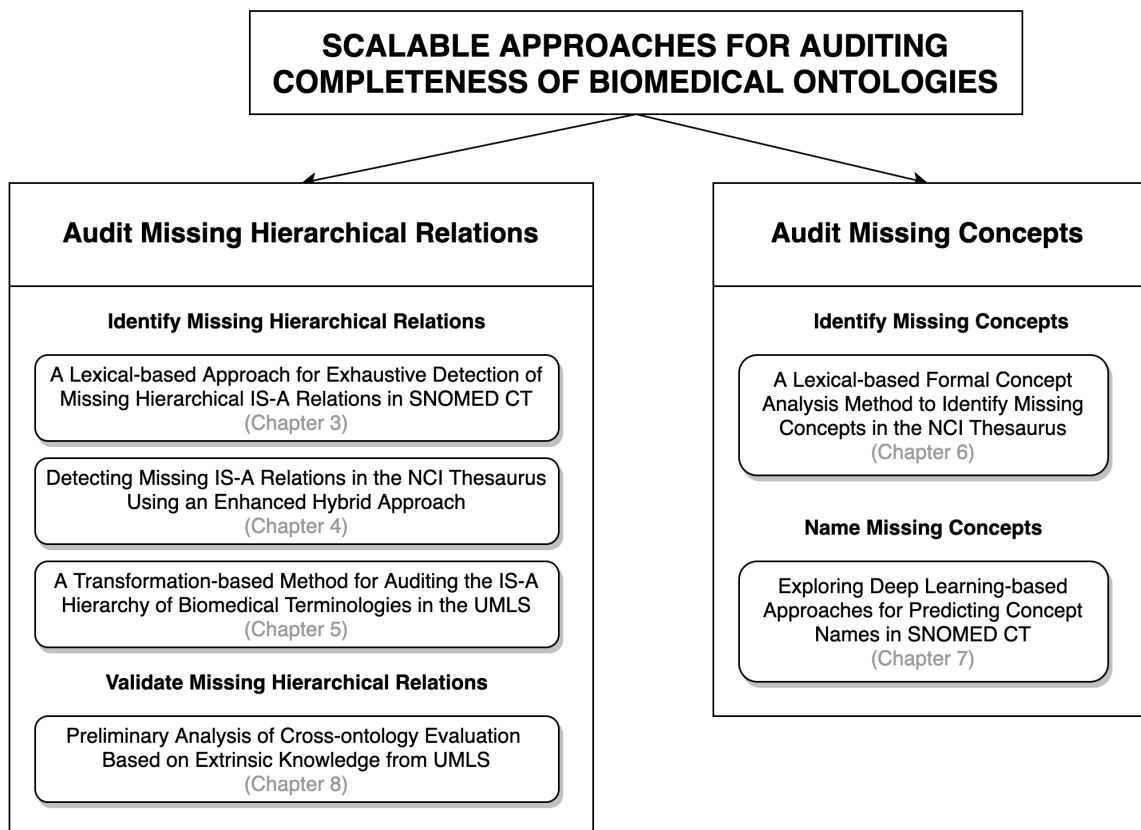
approaches are often limited to sub-hierarchies or part of the hierarchies, and do not scale to the entire ontology. Another limitation of many existing approaches is that they either suggest potential missing concepts or relations purely based on extrinsic knowledge source neglecting the sophisticated intrinsic knowledge [17, 33, 34]; or only indicate potential areas of quality issues based on intrinsic knowledge and then require domain experts’ manual review to come up with the actual issues and solutions, which is time-consuming and labor-intensive [35, 36].

In this dissertation, we develop automated and scalable approaches for detecting potential incompleteness issues, validating the suggested incompleteness issues and providing remediations. Both intrinsic and extrinsic knowledge are utilized. Particularly, the dissertation introduces three approaches in detecting missing hierarchical relations, a Formal Concept Analysis (FCA)-based method to detect missing concepts, a deep learning-based approach to predict concept names and a validation method in which extrinsic knowledge is adopted to collect supporting evidence for the uncovered incompleteness issues.

This dissertation interpolates material from six papers first authored by the author [37–42] and one paper co-authored [28]. Chapter 3 uses material from Reference [37]. Chapter 4 uses material from References [38, 39]. Chapter 5 uses material from Reference [40]. Material from Reference [41] is used in Chapter 6. Material from Reference [42] is used in Chapter 7. Chapter 8 uses material from Reference [28].

An overview of the dissertation is shown in Figure 1.1. In the following, the contribution of each component is discussed.

The semantic meanings of concepts determine if there should be any subsumption relations among concepts. Lexical-based methods have shown great potential in detecting missing hierarchical relations [25, 26, 43]. However, most of them adopt the bag-of-words model, which purely uses single words appearing in the concept name to represent the meanings of concepts. In the first work, we introduce the use of a



**Figure 1.1:** An overview of the auditing approaches introduced in the dissertation.

linguistic feature “noun phrase” which groups a modifier and its corresponding noun as a single feature. It helps reduce the false positives caused by the scenarios when an adjective modifies different nouns in the subconcept and superconcept, which can not be handled by the traditional bag-of-words model.

Besides lexical features, most biomedical ontologies nowadays provide formally defined logical definitions (or role definitions) that refine the meanings of concepts. However, logical definitions are sometimes incomplete, making them impractical to be solely used in representing semantic meanings. In the second hybrid approach applied to the NCI Thesaurus, to model each concept, we combine its lexical features and logical definitions to provide a more comprehensive semantic model. Then a two-step subsumption testing is performed to automatically suggest potentially missing hierarchical relations.

To identify missing hierarchical relations, concepts are usually represented using intrinsic knowledge such as concept names and logical definitions (e.g., in the format of embeddings or feature sets). The effectiveness of these methods to some extent relies on the integrity of the ontology itself (e.g., requires well-defined logical definitions [38]). In the third work, we develop a method based on concept name transformation which also utilizes extrinsic knowledge to identify missing hierarchical relations in the Unified Medical Language System (UMLS) source ontologies. Given a concept, we replace the noun chunks in its name with more general terms to constitute new concept names that are supposed to be more general than the original one. If a new concept name coincides with the name of an existing concept, a hierarchical relation should be established between the two concepts corresponding to the original and new concept names. During this process, knowledge from the audited ontology as well as external ontologies in the UMLS (i.e., both intrinsic and extrinsic knowledge) are leveraged to provide replacement candidates. This results in newly identified missing hierarchical relations that would not be uncovered by only looking into one or two individual ontologies. Also, compared with previous work that usually audit one ontology at a time, this method enables the auditing of multiple source ontologies at the same time.

When it comes to concept enrichment, we propose a lexical-based FCA method to identify missing concepts in the NCI Thesaurus. Existing FCA-based methods [18, 19] mainly utilize logical definitions to search for new or missing concepts. However, the missing concepts they identified only involve logical definitions and no clues are provided on how to name the missing concepts. Therefore, the validation process is laborious. To relieve this, we consider lexical features – words appearing in the concept names as FCA attributes to construct the formal context. In this case, formalizing new concepts also provides bags of words that can be used to name the concepts which are more convenient to validate compared with sets of logical

definitions.

Given bags of words, we further explore deep learning-based approaches to automatically predict concept names that comply with the naming convention of a specific ontology. The task is completed in two parts. In the first part, we adopt three types of neural networks including a simple neural network, a Long Short-Term Memory (LSTM) neural network and a combination of convolutional neural network (CNN) and LSTM as binary classifiers to determine if a given sequence of words is valid or not. In the second part, given a bag of words, pre-trained models are utilized to predict the valid sequence by classifying and filtering all the possible permutations. To the best of our knowledge, this is the first work that automatically predicts names for new or missing concepts in biomedical ontologies.

In general, after identifying incompleteness issues, the results will be manually reviewed by domain experts to evaluate the effectiveness of an auditing method. To relieve the manual burden, in this dissertation, we also present a work that shows the possibility of leveraging extrinsic knowledge (e.g., external ontologies in the UMLS) to validate the detected incompleteness issues. Concepts from the SNOMED CT, NCI Thesaurus and Gene Ontology are mapped to Concept Unique Identifiers (CUIs) in the UMLS and hierarchical relations from multiple source ontologies construct transitive paths to provide supporting evidence for the uncovered missing hierarchical relations.

Several evaluation approaches were adopted to validate the effectiveness of our auditing methods. Random samples (or all the result) of potential incompleteness issues obtained by the three approaches for auditing hierarchical relations and the method for naming missing concepts have been reviewed by domain experts. The performance of each auditing approach is reported in terms of precision, and the performance of the deep learning-based approach is reported in precision, recall and F1-score. The FCA-based concept enrichment method is evaluated by checking whether the new

concepts (i.e., bags of words) are included in any external ontologies in the UMLS.

The incompleteness issues uncovered in this dissertation will be handed over to respective ontology curators so that where appropriate, they could be incorporated into the respective ontologies. Note that the valid results from the hybrid approach have already been added into the newer versions of the NCI Thesaurus.

### **1.3 Organization**

The remainder of this dissertation is organized as follows. Chapter 2 introduces some background about the biomedical ontologies we audited, materials we used and related work in auditing biomedical ontologies. Chapter 3 presents a lexical-based approach to exhaustively detect potentially missing hierarchical relations in the SNOMED CT. Chapter 4 discusses a hybrid approach that combines lexical features and role definitions of concepts to identify potentially missing hierarchical relations within non-lattice subgraphs in the NCI Thesaurus. Chapter 5 introduces a concept name transformation-based method that leverages the UMLS knowledge to identify potentially missing hierarchical relations in its source ontologies. Chapter 6 introduces a lexical- and FCA-based method to identify potentially missing concepts in the NCI Thesaurus. Chapter 7 shows several deep learning models that can be utilized to predict concept names for the missing concepts based on bags of words. Chapter 8 presents a work in which cross-ontology verification based on extrinsic knowledge from the UMLS is used to validate the missing hierarchical relations automatically. Chapter 9 concludes this dissertation and discusses future research directions.

## CHAPTER 2. Background

This chapter first introduces the biomedical ontologies audited in this dissertation, including the SNOMED CT, NCI Thesaurus and Gene Ontology. Computational techniques and an external knowledge source used for supporting quality assurance are then briefly introduced. In addition, this chapter reviews some previous related work in quality assurance of biomedical ontologies.

### 2.1 Biomedical ontologies

Modern ontologies were developed in Artificial Intelligence (AI) to facilitate knowledge representation, knowledge management and knowledge sharing [44]. In biomedicine, biomedical ontologies serve as the semantic scaffolding for us to fully capitalize on the transformative opportunities of the increasingly large amount of digital data produced by the biomedical research enterprise. For example, the BioPortal [45–47], the world’s most comprehensive repository of biomedical ontologies, contains 835 ontologies and over 9 million concepts that have been used to support a wide spectrum of scientific projects.

The principal components of a biomedical ontology are concepts and relations. A concept represents a class of entities within a domain and a relation describes the interaction between concepts. To constrain the interpretation and well-formed use of concepts, biomedical ontologies usually provide human-readable text describing their meanings (e.g., preferred names and synonyms), as well as formally defined properties, features and relations that entities belonging to a concept must have (i.e., logical definitions) [48]. In most cases, the curators of biomedical ontologies state an initial collection of logical definitions for concepts on which DL reasoners such as ELK [49] and Snorocket [50] will further perform reasoning to ensure the consistency of ontologies and achieve more profound inferred definitions (e.g., inferring

subclass/hierarchical relations between concepts based on their stated logical definitions). The inferred hierarchical relations altogether form an inferred hierarchy on which quality assurance (e.g., identifying missing concepts and missing hierarchical relations between concepts) is usually performed.

In this dissertation, we focus on auditing the completeness of three leading biomedical ontologies: SNOMED CT, NCI Thesaurus and Gene Ontology. We will mainly use SNOMED CT community’s nomenclature for further explaining ontological modeling activities and components.

### **2.1.1 SNOMED CT**

SNOMED CT is the most comprehensive, multilingual clinical healthcare terminology in the world. It contains more than 300,000 concepts connected by over 1.5 million relations. SNOMED CT has a broad coverage of health-related topics and it organizes clinical meanings (concepts) into 19 sub-hierarchies, including *Clinical finding*, *Procedure*, *Body structure*, etc [51].

#### **2.1.1.1 Logical model**

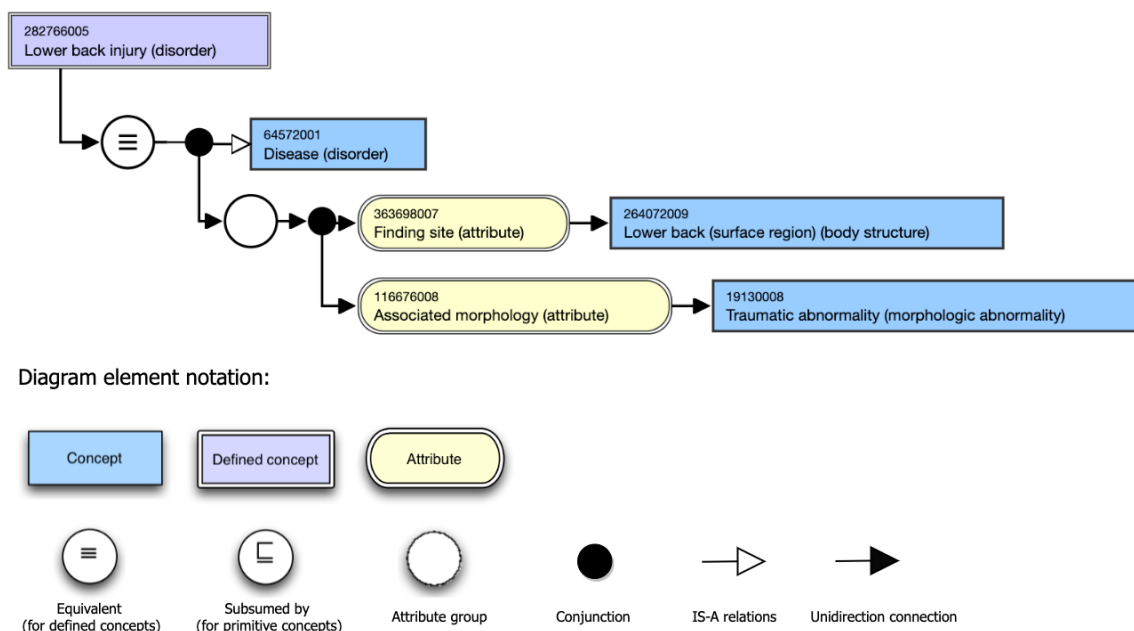
The SNOMED CT logical model specifies a standardized representation of the concepts, the descriptions of concepts, and the relations between concepts.

Each concept in the SNOMED CT represents a unique clinical meaning and has a unique concept identifier (e.g., *282765009*). A fully specified name (FSN) is assigned to each concept providing a unique, unambiguous description of the concept. For instance, concept *282765009*’s FSN is “*Upper back injury (disorder)*” with a semantic tag “*disorder*” in parentheses at the end. Synonyms are also provided for each concept. For example, “*Lumbar region injury*” is a synonym of “*Upper back injury (disorder)*.”

A relation represents an association between two concepts. There are mainly two



types of relations: subtype (or IS-A) relations and attribute relations. The meaning of a concept can be logically defined using subtype and attribute relations. When a concept's definition comprises a number of defining relations, SNOMED CT will put them into different groups to avoid ambiguity as to how they apply. For example, Figure 2.1 shows the stated logical definition of concept “*Upper back injury*,” consisting of one subtype relation (*Is a: Disease*) and a group of two attribute relations ((*Associated morphology: Traumatic abnormality*), (*Finding site: Upper back (surface region)*)). Each single defining relation can be considered as an attribute-value pair (attribute: value) for the source concept. Each group of defining relations are treated as an integration, thus a group of attribute-value pairs. In this example, “*Upper back injury*” is considered as the “*traumatic abnormality*” that happened at “*upper back (surface region)*.” In the SNOMED CT, there are over 50 attributes that can be used as the attribute “type” of a defining relation, including “*Is a*,” “*Finding site*,” “*Causative agent*,” “*Clinical course*,” and “*Laterality*.” And the value of an attribute is another concept in the SNOMED CT.



**Figure 2.1:** Diagram that shows the **stated** logical definition of concept *Upper back injury (disorder)* (282765009) [52] and notations of diagram elements [53].

Besides constructing the definitions of concepts, IS-A relations in the SNOMED CT also organize concepts into a hierarchy. Thus, IS-A relations are also taken as hierarchical relations. The hierarchy organizes the concepts from the more general ones to the more detailed (or specific) ones. The most general concept in the SNOMED CT is “*SNOMED CT Concept*” (138875005). Concepts that are more general are usually placed at the top part of the hierarchy and then at each level down the hierarchy, the concepts become increasingly more specific. In other words, a concept is more detailed than its supertype and is more general than its subtype. In the previous example, “*Upper back injury*” is a subtype of “*Disease*.” Therefore, the definition of “*Disease*” is more general than that of its subtype “*Upper back injury*.”

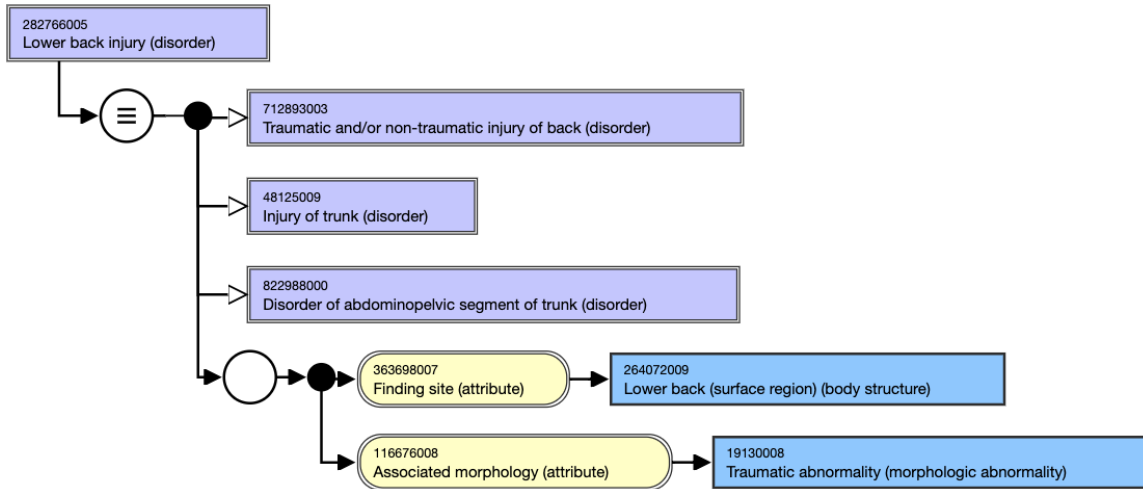
The status of a concept’s definition in the SNOMED CT is either *sufficiently/fully defined* or *primitive*. A concept is considered to be sufficient defined if its definition is sufficiently to distinguish its meaning from other similar concepts. Otherwise, the definition status of a concept is primitive. The definition status is usually decided by the curators of the SNOMED CT. When a concept is sufficiently defined, any concept whose definition satisfies its defining relations (i.e., whose definition is more detailed than its) can be considered as being equivalent to or a subtype of the concept. For example, concept “*Upper back injury*” is sufficiently defined and therefore, any concepts satisfying the defining relations in Figure 2.1 are either equivalent to or subtypes of “*Upper back injury*.”

#### **2.1.1.2 Stated and inferred logical definitions**

The stated definition of a concept in SNOMED CT is a set of relations (grouped or ungrouped) that the ontology curator has stated to be defining characteristics of the concept. For instance, the stated definition of concept “*Upper back injury*” consists of a subtype relation (*Is a: Disease*) as well as a group of two attribute relations (*Associated morphology: Traumatic abnormality*) and (*Finding site: Upper*

*back (surface region)).*

Inferred concept definitions are sets of non-redundant (most specific) defining relations derived by the DL reasoners. DL reasoners can check the consistency of stated relations across the whole ontology and infer a hierarchy of concepts based on the stated facts (i.e., infer new hierarchical relations). Figure 2.2 shows the inferred definition of concept “*Upper back injury.*” Note that (*Is a: Disease*) is not included in the inferred definition because it is redundant to (i.e., more general than) other newly inferred subtype relations, such as (*Is a: Injury of trunk*).



**Figure 2.2:** Diagram that shows the **inferred** logical definition of concept *Upper back injury (disorder)* (282765009) [52].

### 2.1.2 National Cancer Institute Thesaurus

NCI Thesaurus covers knowledge across a wide range of cancer research domains, including cancer-related diseases, findings and abnormalities; genes and gene products; therapies, drugs and chemicals, etc. It combines terminologies from numerous cancer research related domains and provides a way to integrate or link these kinds of information together through semantic relations. It has been used in a growing number of NCI and other systems to facilitate data sharing and interoperability [54, 55].

The 20.12d version of the NCI Thesaurus contains over 150,000 concepts, 120,000

textual definitions and more than 40,000 relations between concepts [56]. Figure 2.3 shows the inferred definitions and associations of concept “*Benign Skin Neoplasm*” (*C2896*) as they are displayed in the NCITerm Browser [3]. The “role relationships” are essentially the same as the attribute relations in the SNOMED CT.

| Benign Skin Neoplasm (Code C2896)   |   |
|---|---|
| Terms & Properties  | Synonym Details                               |
| Relationships   | Mappings                                      |
| View All  |   |
| <b>Relationships with other NCI Thesaurus Concepts</b>  |   |
| Parent Concepts:  |   |
| <a href="#">Benign Neoplasm</a>   |   |
| <a href="#">Skin Neoplasm</a>   |   |
| Child Concepts:   |   |
| <a href="#">Adult Xanthogranuloma</a>   |   |
| <a href="#">Benign Dermal Neoplasm</a>  |   |
| <a href="#">Benign Epithelial Skin Neoplasm</a>   |   |
| <a href="#">Benign Scrotal Neoplasm</a>   |   |
| <a href="#">Benign Skin Appendage Neoplasm</a>  |   |
| <a href="#">Benign Skin Melanocytic Nevus</a>   |   |
| <a href="#">Cutaneous Ganglioneuroma</a>  |   |
| <a href="#">Dermoid Cyst of the Skin</a>  |   |
| <a href="#">Juvenile Xanthogranuloma</a>  |   |
| Role Relationships, asserted or inherited, pointing from the current concept to other concepts:<br>(True for the current concept and its descendants, may be inherited from parent(s).) |   |
| Relationship  | Value (qualifiers indented underneath)        |
| Abnormal Cell   |   |
| Disease_Excludes_Abnormal_Cell  | <a href="#">Malignant Cell</a>                |
| Disease_Has_Abnormal_Cell   | <a href="#">Neoplastic Cell</a>               |
| Anatomic Structure, System, or Substance  |   |
| Disease_Has_Associated_Anatomic_Site  | <a href="#">Integumentary System</a>          |
| Disease_Has_Associated_Anatomic_Site  | <a href="#">Skin</a>                          |
| Disease_Has_Primary_Anatomic_Site   | <a href="#">Skin</a>                          |
| Disease, Disorder or Finding  |   |
| Disease_Excludes_Finding  | <a href="#">Dysplasia</a>                     |
| Disease_Excludes_Finding  | <a href="#">Invasive Lesion</a>               |
| Disease_Excludes_Finding  | <a href="#">Malignant Cellular Infiltrate</a> |
| Disease_Has_Finding   | <a href="#">Benign Cellular Infiltrate</a>    |
| Disease_Has_Finding   | <a href="#">Cutaneous Involvement</a>         |
| Associations pointing from the current concept to other concepts:<br>(True for the current concept.)  |   |
| Relationship  | Value (qualifiers indented underneath)        |
| Concept_In_Subset   | <a href="#">CPTAC CodeLists Terminology</a>   |
| Concept_In_Subset   | <a href="#">CPTAC Neoplasms Codelist</a>      |
| Concept_In_Subset   | <a href="#">CPTAC Terminology</a>             |

**Figure 2.3:** Inferred defining relations and associations of concept *Benign Skin Neoplasm* (*C2896*) in the NCI Thesaurus [3].

### 2.1.3 Gene Ontology

Developed to address the need for a consistent description of gene products in different databases, Gene Ontology (GO) strives to provide unified representations of genes, gene products and sequences [57, 58]. Gene Ontology maintains knowledge in three main aspects: molecular function, biochemical activity of gene product; biological

process that is the biological objective to which the gene or gene product contributes; and cellular component referring to the place in the cell where a gene product is active [59, 60].

The 01/01/2021 release of Gene Ontology provides over 44,000 GO Terms (i.e., concepts) and over 7,000,000 annotations (i.e., relations) [61].

## 2.2 Techniques and extrinsic knowledge for supporting auditing

In this dissertation, Formal Concept Analysis (FCA) is utilized to identify potentially missing concepts and deep learning techniques are adopt to predict names for the missing concepts. Extrinsic knowledge from the UMLS is utilized to detect potentially missing hierarchical relations and to provide supporting evidence for the uncovered incompleteness issues.

### 2.2.1 Formal Concept Analysis

Formal Concept Analysis (FCA) is a mathematical theory concerned with the formalization of concepts and conceptual thinking [62]. With FCA, we can generate a concept hierarchy from a collection of objects and attributes. The input of FCA is *formal context*  $K = (O, A, R)$ , where  $O$  is a set of objects,  $A$  is a set of attributes, and  $R$  is a binary relation between  $O$  and  $A$ . The notation  $(o, a) \in R$  means that object  $o$  has attribute  $a$ .

Each formal context  $K$  induces two operators: derivation operators  $\uparrow: 2^O \rightarrow 2^A$  and concept-forming operators  $\downarrow: 2^A \rightarrow 2^O$ . The operators are defined, for each  $X \subseteq O$  and  $Y \subseteq A$ , as follows:

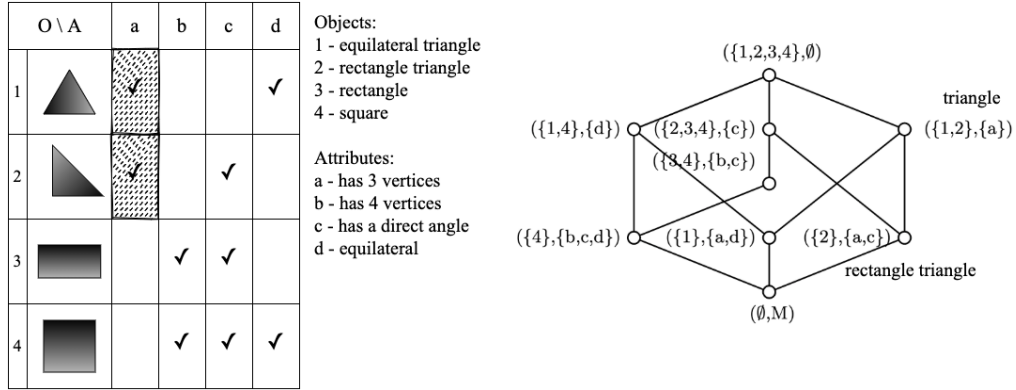
$$X^\uparrow = \{a \in A \mid \forall o \in X: (o, a) \in R\},$$

$$Y^\downarrow = \{o \in O \mid \forall a \in Y: (o, a) \in R\},$$

where  $X^\uparrow$  is the set of all attributes shared by all objects in  $X$ , and  $Y^\downarrow$  is the set of all objects sharing all attributes in  $Y$ .

A formal concept of  $K$  is a pair  $(X, Y)$  with  $X \subseteq O$  and  $Y \subseteq A$  such that  $X^\uparrow = Y$  and  $Y^\downarrow = X$ . The subconcept-superconcept relation between formal concepts is given by  $(X_1, Y_1) \leq (X_2, Y_2)$  iff  $X_1 \subseteq X_2$  ( $Y_2 \subseteq Y_1$ ). All formal concepts derived from the formal context  $K$  together with the subconcept-superconcept relations form a complete lattice [63]. Note that lattice is a desired property for well-structured ontologies [64].

Figure 2.4 shows a simple formal context (i.e., the table on the left) constructed by four geometry figures and four attributes. In this example, formal concept  $(\{1,2\}, \{a\})$  (i.e., triangle) could be derived since object equilateral triangle and rectangle triangle share one attribute *has 3 vertices*; and inversely, attribute *has 3 vertices* is only shared by these two objects. The graph on the right shows the complete lattice formed by all formal concepts and subconcept-superconcept relations derived from the given formal context.



**Figure 2.4:** Example formal context and the complete lattice formed by all formal concepts derived [62]. The formal context is of four geometry figures and four attributes. Formal concept  $(\{1,2\}, \{a\})$  is marked by dot in the formal context.

### 2.2.2 Long Short-Term Memory (LSTM)

Deep learning approaches have been widely used in text-related tasks such as language modeling, textual analysis, information retrieval and sequence generation, and showed better performance than traditional machine learning methods in these tasks [65]. Among different types of neural networks, LSTM [66] is proven to be suitable for various sequential tasks such as speech recognition, translation and protein secondary structure prediction. This is basically because LSTM is able to capture long-term temporal dependencies efficiently [67]. Many previous works adopted LSTM-based methods for sequence classification and sequence generation. Graves et al. introduced bidirectional LSTM (BLSTM) which could learn dependencies both forwardly and backwardly [68]. He et al. adopt bidirectional LSTM to distinguish varied meanings of ambiguous biomedical terms in biomedical texts [69]. Zhou et al. combined CNN and recurrent neural network (RNN) and proposed a unified model called C-LSTM for sentence representation and text classification [70].

### 2.2.3 Unified Medical Language System

The UMLS, developed by the US National Library of Medicine, integrates various health and biomedical vocabularies and standards to enable interoperability between different applications and systems. It has been used in supporting a wide range of applications in biomedicine including information retrieval, natural language processing (NLP), deep learning, phenotyping, and clinical decision support [71–83]. The UMLS consists of three knowledge sources: the Metathesaurus that contains concepts from many ontologies, the Semantic Network that contains semantic types and their relationships, and the SPECIALIST Lexicon and Lexical Tools to facilitate NLP [84–88].

The UMLS Metathesaurus is organized by concept or meaning. Since a concept can have many different names, the UMLS Metathesaurus links all the names from different source ontologies that have the same meaning. Every occurrence of a concept

name (or string) in each source ontology is the basic building block or “atom” of the UMLS Metathesaurus and assigned a unique atom identifier (AUI). Atoms with the same meaning are mapped to a concept assigned a concept unique identifier (CUI). For example, consider a concept in the SNOMED CT with ID *282766005* and preferred name “*Lower back injury.*” It also has a synonym “*Lumbar region injury [52].*” In the UMLS Metathesaurus, the AUI for its preferred name is *A3255024* and the AUI for its synonym is *A3288211*. These two atoms are both mapped to the same UMLS concept with CUI *C0560632*, which has a total of 14 atoms mapped from different source ontologies. The UMLS preserves the relations between concepts from its source ontologies. For instance, the IS-A relation between the atom “*Superficial injury of lower back*” with AUI *A28900983* and the atom “*Lower back injury*” with AUI *A3255024* comes from SNOMED CT.

In addition, each UMLS concept (CUI) is assigned at least one semantic type in order to provide a consistent categorization of all concepts. For example, the concept “*Lower back injury*” (CUI: *C0560632*) is assigned a semantic type “*Injury or Poisoning.*” There are currently 127 semantic types in the UMLS such as “*Disease or Syndrome,*” and “*Therapeutic or Preventive Procedure.*”

## 2.3 Quality assurance of biomedical ontologies

In this section, previous related work in auditing concepts and hierarchical relations in biomedical ontologies is reviewed.

### 2.3.1 Auditing concepts

When it comes to quality assurance of concepts, completeness and consistency are two properties that gained a lot of awareness.



### 2.3.1.1 Auditing concept completeness

There are two types of approaches to identify missing or new concepts for the concept enrichment of biomedical ontologies.

The first type mainly leverages extrinsic knowledge (e.g., imports concepts from external sources). For instance, Chandar et al. developed a similarity-based method that suggested extracted phrases from text corpus as new concepts for the SNOMED CT [20]. Peng et al. analyzed connected matrices from Gene Ontology and biological network to identify new terms for Gene Ontology [21]. He et al. leveraged alignments between different ontologies to suggest new concepts for the SNOMED CT [16] and NCI Thesaurus [17]. For these methods, the sophisticated intrinsic knowledge to some extent is neglected.

The other type mainly utilizes the intrinsic knowledge within the ontology itself. Jiang and Chute performed Formal Concept Analysis (FCA) based on logical definitions to search for possible missing concepts in the SNOMED CT [18]. However, due to the computational limitation, their method was only applied to a small portion of SNOMED CT concepts. Zhu et al. improved Jiang and Chute’s work by developing a scalable multistage algorithm called Spark-MCA [19] that enabled an exhaustive FCA evaluation on all the SNOMED CT concepts. However, a limitation of these two FCA-based approaches is that the potentially missing concepts identified only involve logical definitions and no concept names were provided. Therefore, it is difficult to validate those missing concepts.

Zhang et al. [64, 89] introduced a lattice-based evaluation of ontologies. Lattice is a desirable property for a well-formed ontology. A pair of concepts is known as a non-lattice pair, if the two concepts have more than one maximal shared descendant. A non-lattice subgraph is obtained from a non-lattice pair by reversely computing the minimal common ancestors of the maximal common descendants of the non-lattice pair and aggregating all the concepts and edges together. Cui et al. introduced a

structural-lexical method by mining lexical patterns in non-lattice subgraphs, where one of the patterns called “Union-Intersection” can automatically identify missing concepts in the SNOMED CT [22].

### **2.3.1.2 Auditing concept modeling consistency**

Besides auditing the completeness of concepts, previous work also paid attention to concept modeling consistency throughout the biomedical ontologies.

Burse et al. proposed a stop-word based contextual auditing method to identify inconsistencies in the modeling of SNOMED CT concepts [90]. They summarized a few patterns that associated stop words (e.g., “and,” “with,” “and/or”) and content surrounding stop words with the appearance of logical definitions. Concepts whose names contain stop words but violate the patterns are signs of inconsistencies.

Verspoor et al. developed a transformation-based clustering methodology to identify terms in Gene Ontology that express similar semantics but use different linguistic conventions [91] (e.g., “X Y” versus “Y of X”). However, much manual effort was required to review the clusters and uncover potential inconsistent conventions.

### **2.3.2 Auditing hierarchical relations**

Existing approaches for auditing hierarchical relations can be roughly classified into the following categories: structural-based, lexical-based, structural-lexical-based, and machine learning-based.

#### **2.3.2.1 Structural-based approaches**

Structural-based approaches mainly rely on the relations between concepts to help reveal potentially missing hierarchical relations.

Abstraction networks (AbNs) [33, 35, 36, 92–95], which group concepts based on shared outgoing relations, have been extensively studied to identify problematic areas in ontologies that may contain quality issues, including missing hierarchical relations.

However, manual review of problematic areas by domain experts is required to locate and uncover the exact quality issues which is labor-intensive.

Chen et al. presented a recursive method to locate missing hierarchical relations in the UMLS Metathesaurus [24]. Concepts were first partitioned into semantically uniform sets based on their semantic types. Then domain experts helped insert smaller clusters into larger clusters during which process missing hierarchical relations could be generated. Still, they require domain expert to go through many possible combinations and come up with the missing hierarchical relations manually.

### **2.3.2.2 Lexical-based approaches**

Lexical-based approaches mainly utilize lexical features or patterns of concepts to identify missing hierarchical relations.

Quesada-Martínez et al. analyzed concept names in the SNOMED CT to identify lexical regularities (LR) and suggested missing relations (including missing hierarchical relations) [29]. However, only a small amount of LR could be used to generate missing relations.

Abeyasinghe et al. introduced a lexical-based inference approach to derive hierarchical inconsistencies and uncover missing hierarchical relations in Gene Ontology by comparing linked and unlinked concepts using words in concept names [27].

Abeyasinghe et al. developed a Subsumption-based Sub-term Inference Framework (SSIF) [96] to detect missing hierarchical relations in Gene Ontology. Concept names were re-modeled based on the part of speech, concept name inclusion relations and antonyms appearing in concept names. Three conditional rules were developed for backward subsumption inference.

Bodenreider considered individual words appearing in concept names (i.e., lexical features) as logical definitions of concepts, and used DL classifier to automatically derive hierarchical relations among concepts in the SNOMED CT; and then compared

the DL-derived hierarchy with the original SNOMED CT hierarchy to detect missing hierarchical relations [25]. One thing is that individual words may not be sufficient or accurate in representing the semantic meaning of a concept.

### **2.3.2.3 Structural-lexical-based approaches**

Structural-lexical-based approaches utilize both concepts’ lexical features and relations between concepts. Previously, we have investigated approaches combining non-lattice subgraphs and lexical features of concepts to automatically suggest missing hierarchical relations in the SNOMED CT [22, 26] and NCI Thesaurus [23, 43]. Basically, the first step was to extract non-lattice subgraphs (i.e., areas with a higher possibility to contain quality issues). Then, either lexical patterns among concepts [22, 23] or enriched lexical features [26, 43] were utilized to identify missing hierarchical relations.

### **2.3.2.4 Machine learning-based approaches**

Besides the rule-based approaches mentioned above, recently, machine learning-based approaches have also been explored to help detect or validate missing hierarchical relations in biomedical ontologies [31, 32, 97].

Sun et al. explored deep learning-based approaches to validate incompleteness detected by non-lattice-based auditing methods [32]. In their work, defining relations of concepts were converted into embeddings and a CNN classifier is developed to predict hierarchical relations for given concept pairs.

Liu et al. generated descriptions for concepts based on their structural information (e.g., parents, children, siblings) and used Doc2Vec to learn vector representations of concepts (or concept embeddings) in the NCI Thesaurus. Then a CNN model was trained to predict if there is any subsumption relations between two given concepts [31]. In another work [97], Liu et al. employed Bidirectional Encoder Representations from Transformers (BERT) to come up with the embeddings for concepts

and used a similar pipeline for subsumption checking.

One thing is that the performance of machine learning-based approaches highly relies on the strategy of selecting positive/negative samples for training. For instance, Sun et al. [32] selected training samples based on a combination of features such as graph similarity, path length to the root, concept name similarity, etc. Liu et al. [31, 97] trained the model using concept pairs from AbNs satisfying certain structural relations (e.g., uncle and its nephews). In these cases, although good performance has been achieved for hierarchical relation classification on the pre-constructed training and testing data in the traditional machine learning manner, the trained model cannot be directly used to uncover missing hierarchical relations due to many false predictions, and always need extra assistance (e.g., AbNs in [31] and non-lattice subgraphs in [32]) to provide candidate pairs of concepts to reduce false predictions.

## CHAPTER 3. A Lexical-based Approach for Exhaustive Detection of Missing Hierarchical IS-A Relations in SNOMED CT

This chapter introduces a lexical-based method to detect missing hierarchical relations in the SNOMED CT. There are mainly three steps. First, a set of stop words and antonym pairs are leveraged to avoid potential erroneous missing hierarchical relations. Then, each concept is modeled as an enriched set of lexical features, which consists of words and noun phrases in the name of the concept and concept’s ancestors. The semantic meaning of a concept is represented by the lexical feature set. At last, subset inclusion checking between lexical feature sets is performed to automatically derive missing hierarchical relations.

### 3.1 Methods

#### 3.1.1 Identifying stop words/phrases and antonym pairs

Stop words/phrases may result in wrongly suggested missing hierarchical relations (or false positives). Take concepts “*Velopharyngeal incompetence due to cleft palate (disorder)*” and “*Cleft palate (disorder)*” as an example, even though the set of lexical features of the former concept contains that of the latter concept, there should not be any subsumption relations between these two concepts. Words/phrases such as “due to” are highly likely to suggest false positives and thus are considered as stop words/phrases. In this work, we leveraged a list of stop words/phrases used in previous work [26], including: “and,” “or,” “no,” “not,” “without,” “due to,” “secondary to,” “except,” “by,” “after,” “co-occurrent,” “bilateral,” “examination,” “able,” “amputation,” “removal,” “replacement,” “resection,” “excision,” “reaction to,” “unable,” “failure,” “failed,” “abnormal,” “excluding,” “non,” and “pre.”

Similarly, concept pairs whose lexical features contain antonym pairs are also likely to generate erroneous suggestions. For instance, considering concepts “*Secondary ma-*

*lignant neoplasm of right upper lobe of lung (disorder)*” and “*Neoplasm of right lower lobe of lung (disorder)*,” apparently there should not be any subsumption relations between these two concepts, since the former concept is related to “right upper lobe of lung” while the latter concept is related to “right lower lobe of lung”. However, if the former concept inherited a lexical feature “lower” from one of its ancestors “*Malignant neoplasm of lower respiratory tract (disorder)*,” then the lexical feature set of the former would subsume that of the latter, as a result of which an incorrect hierarchical relation between the former and latter would be suggested. To collect such potential antonym pairs, we adopted a list of adjective antonym pairs from WordNet [98], including (“open,” “closed”), (“acute,” “chronic”), (“right,” “left”), etc. We also identified additional antonym pairs that are not included in WordNet, such as (“upper,” “lower”).

### 3.1.2 Constructing lexical features for concepts

Most existing lexical-based methods for the identification of missing hierarchical relations use words in concept names as the lexical features of concepts. In this work, we model concepts not only using words, but also utilizing noun phrases. For each concept, we first preprocess its fully specified name (FSN) and identify an initial set of lexical features consisting of words and noun phrases in the concept’s FSN. Then we enrich the set with more lexical features inherited from the concept’s ancestors.

#### 3.1.2.1 Preprocessing FSNs of concepts

We preprocess the FSNs of concepts before the initialization of lexical feature sets. For each concept, we split its FSN (by space) into words sequentially and remove its semantic tag (e.g., “(disorder)”). The semantic tag will be leveraged while suggesting potentially missing hierarchical relations. We further process special symbols in FSNs such as removing parentheses and square brackets, and replacing

backslash with “or” if the FSN does not contain numbers (e.g., “Sickness/injury care” will result in “Sickness or injury care,” while “5 mg/ml” will remain intact).

### 3.1.2.2 Initializing lexical feature sets with noun phrases and words

In this work, instead of purely using the bag-of-words model, we consider noun phrases as meaning features of concepts to facilitate the identification of missing hierarchical relations. Take two concepts “*Acute sensitivity to pain (finding)*” and “*Acute pain (finding)*” as an example, if we simply used the bag-of-words model, their lexical feature sets would be {acute, sensitivity, to, pain} and {acute, pain} respectively, where {acute, sensitivity, to, pain} is a superset of {acute, pain} (i.e., more detailed), and thus “*Acute sensitivity to pain (finding)*” would be suggested as a subtype of “*Acute pain (finding)*.” However, this suggestion is incorrect since “*Acute sensitivity to pain*” is a finding of pain threshold, while “*Acute pain*” is a finding of pattern of pain; and there should not be any subsumption relations between these two concepts. The reason for this incorrect suggestion is that the adjective “acute” is the modifier for two different nouns (“sensitivity” and “pain”) in these two concepts. To avoid such situation, we model a concept’s name as a set of noun phrases and words, where a noun phrase groups the modifier(s) and the corresponding noun as a single feature. Hence in the above example, the two concepts’ lexical features will become {acute, sensitivity, to, pain, acute sensitivity} and {acute, pain, acute pain}, which do not have any subset-superset relation anymore.

We use Stanford CoreNLP Parser [99] to identify noun phrases. Note that the parser may recognize noun phrases in different levels of granularity. For instance, for concept “*Anesthesia for procedure on veins of lower leg (procedure)*,” there is a base level noun phrase “lower leg” which is a component of a higher level noun phrase “veins of lower leg.” In this work, we only consider the base level noun phrases. That is, we model a concept’s FSN initially as a set of individual word(s) and base level



noun phrase(s). In this example, the initial set of lexical features for the concept is {anesthesia, for, procedure, on, veins, of, lower, leg, lower leg}.

### 3.1.2.3 Enriching lexical feature sets

We enrich concepts' lexical features in two steps. In the first step, for each concept  $c$ , we check if its FSN contains noun phrase(s) identified in the initial feature sets of other concepts that are not hierarchically linked with  $c$ ; and if yes, we add such noun phrase(s) into  $c$ 's initial lexical feature set. We denote concept  $c$ 's set of lexical features obtained after the first-step enrichment process as  $E_{1c}$ . In the second step, for each concept, we further enrich its set of lexical features with its ancestors' sets of lexical features. It is intuitive that if concept  $x$  is a subtype of concept  $y$ , then the lexical features or attributes of concept  $y$  are also considered to be true for concept  $x$  (i.e.,  $x$  inherits  $y$ 's attributes). In this work, we maintain a directed graph that is constructed using all the inferred hierarchical relations in the SNOMED CT, compute its transitive closure, and obtain the ancestors of concepts using breadth-first search (BFS). While performing the second-step enrichment process for a concept  $c$ , if an ancestor  $a$  contains stop word(s)/phrase(s), then we do not add  $a$ 's set of lexical features to  $c$ 's. More formally, we have

$$E_{2c} = E_{1c} \cup \left( \bigcup \{E_{1a} \mid a \in A_c \text{ and } a \text{ does not contain any stop words/phrases}\} \right),$$

where  $E_{2c}$  denotes concept  $c$ 's set of lexical features after the second-step enrichment process, and  $A_c$  is the set of  $c$ 's ancestors.

Table 3.1 shows an example of the initial and enriched sets of lexical features for concept  $c$ : 371977004 – “*Primary malignant neoplasm of cecum (disorder)*.” The noun phrase identified in the initial set of lexical features is “primary malignant neoplasm” (underlined). After the first-step enrichment ( $E_{1c}$ ), a new noun phrase “malignant neoplasm” is identified from concepts which are not hierarchically linked with  $c$ . After

the second-step enrichment ( $E_{2c}$ ), more noun phrases and words are inherited from  $c$ 's ancestors. For instance, noun phrase "large intestine" is inherited from the initial lexical feature set of  $c$ 's parent – "*Primary malignant neoplasm of large intestine (disorder)*" and noun phrase "malignant tumor" is inherited from  $c$ 's other parent – "*Malignant tumor of cecum (disorder)*."

**Table 3.1:** The initial and enriched sets of lexical features of an example concept  $c$ : 371977004 – *Primary malignant neoplasm of cecum (disorder)*. Noun phrases are underlined.

|                       |  |
|-----------------------|--|
| $c$ 's FSN            | Primary malignant neoplasm of cecum (disorder)   |
| Initial set           | {primary, malignant, neoplasm, of, cecum, <u>primary malignant neoplasm</u> }  |
| Enriched set $E_{1c}$ | {primary, malignant, neoplasm, of, cecum, <u>primary malignant neoplasm</u> , <u>malignant neoplasm</u> }  |
| Enriched set $E_{2c}$ | {primary, malignant, neoplasm, of, cecum, <u>primary malignant neoplasm</u> , <u>malignant neoplasm</u> , abdominal, mass, <u>abdominal mass</u> , disorder, digestive, structure, <u>digestive structure</u> , finding, large, intestine, large intestine, neoplastic, disease, <u>neoplastic disease</u> , malignant neoplastic disease, viscus, <u>structure finding</u> , body, region, body region, trunk, <u>trunk structure</u> , abdomen, tumor, malignant tumor, organ, digestive organ, gastrointestinal, tract, gastrointestinal tract, system, digestive system, intraabdominal, intraabdominal organ, bowel, bowel finding, lower, <u>lower gastrointestinal tract</u> , <u>body system</u> , abdominal organ finding, <u>abdominal organ</u> , <u>gastrointestinal tract finding</u> , segment, <u>abdominal segment</u> , intestinal, <u>intestinal tract</u> , <u>digestive system finding</u> , <u>system finding</u> , body structure, digestive tract, cecal, <u>cecal mass</u> } |

### 3.1.3 Identifying potentially missing hierarchical relations

To automatically suggest potentially missing hierarchical relations, we first produce candidate pairs of concepts (say  $x$  and  $y$ ) that meet the following conditions:

- concepts  $x$  and  $y$  are within the same sub-hierarchy (we assume that concepts in different sub-hierarchies do not have hierarchical relations since sub-hierarchies in SNOMED CT do not share common concepts);
- $x$  and  $y$  are not hierarchically linked through existing hierarchical relations;

- $x$  and  $y$  share the same semantic tag;
- neither  $x$  nor  $y$  contains any stop word/phrase; and
- the enriched sets of lexical features  $E_{2x}$  and  $E_{2y}$  do not contain antonym pairs.

Then for each candidate pair of concepts  $(x, y)$ , we systematically compare their enriched sets of lexical features  $E_{2x}$  and  $E_{2y}$  as follows: if  $E_{2x}$  is a superset of  $E_{2y}$ , then “concept  $x$  IS-A concept  $y$ ” will be suggested as a potentially missing hierarchical relation; if  $E_{2x}$  is a subset of  $E_{2y}$ , then “concept  $y$  IS-A concept  $x$ ” will be suggested as a potentially missing hierarchical relation; otherwise, nothing will be suggested.

Since our suggestion of missing hierarchical relations is in an exhaustive way, it may result in redundant missing hierarchical relations. For example, our approach may suggest “ $A$  IS-A  $B$ ” and “ $A$  IS-A  $C$ ” as two missing hierarchical relations, while  $B$  is an ancestor of  $C$  in the current ontology (i.e., an existing IS-A relation). In this case, “ $A$  IS-A  $B$ ” is considered redundant because it can be implied by the potentially missing hierarchical relation “ $A$  IS-A  $C$ ” and the existing relation “ $C$  IS-A  $B$ ”. To improve the evaluation efficiency, we avoid unnecessary analyses on such redundant relations. More formally, a detected potentially missing hierarchical relation “ $A$  IS-A  $B$ ” is considered as redundant if it can be inferred by other missing or existing hierarchical relations.

### 3.2 Results

We applied our method to all the sub-hierarchies in the September 2017 release of SNOMED CT (US edition), except “*SNOMED CT Model Component (metadata)*” (e.g., definition status) and “*Special concept (special concept)*” (e.g., inactive concept). A total of 38,615 potentially missing hierarchical relations were suggested. Table 3.2 shows the number of potentially missing hierarchical relations in each sub-hierarchy. For instance, 6,946 potentially missing hierarchical relations were identified from the “*Clinical finding*” sub-hierarchy.

**Table 3.2:** Numbers of missing hierarchical relations detected in terms of the sub-hierarchies.

| Sub-hierarchy     | # of Potentially Missing IS-A | Sub-hierarchy                        | # of Potentially Missing IS-A |
|-------------------|-------------------------------|--------------------------------------|-------------------------------|
| Body structure    | 26,161                        | Situation with explicit context      | 82                            |
| Clinical finding  | 6,946                         | Staging and scales                   | 36                            |
| Procedure         | 3,861                         | Social context                       | 33                            |
| Substance         | 390                           | Specimen                             | 31                            |
| Organism          | 277                           | Environment or geographical location | 26                            |
| Observable entity | 242                           | Pharmaceutical / biologic product    | 22                            |
| Physical object   | 234                           | Record artifact                      | 2                             |
| Qualifier value   | 185                           | Physical force                       | 0                             |
| Event             | 87                            |                                      |                               |

### 3.2.1 Evaluation

To evaluate the effectiveness of our approach for detecting missing hierarchical relations, we randomly selected a sample of 100 potentially missing hierarchical relations from the “*Clinical finding*” sub-hierarchy. A domain expert reviewed the sample and verified that 90 out of 100 missing hierarchical relations are valid (or true positives), indicating that our approach achieved a precision of 90%. Table 3.3 lists 15 examples of missing hierarchical relations in the “*Clinical finding*” sub-hierarchy verified by the domain expert, including “*Open injury of diaphragm (disorder)*” IS-A “*Open wound of thorax (disorder)*,” and “*Primary malignant neoplasm of fibula (disorder)*” IS-A “*Malignant neoplasm of long bone of lower leg (disorder)*.”

For each false positive (i.e., invalid missing hierarchical relation suggested), we provided the domain expert with the existing hierarchical relation(s) which lead to the suggestion of the false positive. The domain expert further reviewed these existing hierarchical relations and checked whether any of them is problematic.

### 3.2.2 Analysis of false positive cases

We manually examined the false positive cases for potential causes. For instance, our approach suggests a false positive: “*Familial malignant neoplasm of pancreas (disorder)*” IS-A “*Malignant tumor of body of pancreas (disorder)*,” since the former concept

inherits a lexical feature “body” from its ancestor “*Mass of body region (finding)*.” However, the meaning of “body” in “*Malignant tumor of body of pancreas (disorder)*” is different than its meaning in “*Mass of body region (finding)*.” The former refers to the finding site of structure of body of pancreas, while the latter refers to the finding site of body region structure. Therefore, there should not be a hierarchical relation between the two concepts. In this case, the false positive is due to the varied meanings of a word in different contexts.

**Table 3.3:** Examples of missing hierarchical relations in the “*Clinical finding*” sub-hierarchy confirmed by the domain expert.

| Subconcept   | Superconcept   |
|--|--|
| Primary adenosquamous cell carcinoma of larynx (disorder)            | Carcinoma of larynx (disorder)                           |
| Strain of fascia of intrinsic muscle of thumb (disorder)             | Injury of fascia of intrinsic muscle of thumb (disorder) |
| Pelvic muscular dystrophy (disorder)                                 | Degenerative disorder of muscle (disorder)               |
| Superficial injury of interscapular region with infection (disorder) | Superficial injury of trunk with infection (disorder)    |
| Contracture of iliopsoas (disorder)                                  | Disorder of soft tissue of trunk (disorder)              |
| Carcinoma in situ of upper labial mucosa (disorder)                  | Tumor of upper labial mucosa (disorder)                  |
| Complete ankylosis of the spine (disorder)                           | Disorder of vertebra (disorder)                          |
| Plasmodium vivax malaria with rupture of spleen (disorder)           | Infectious disease of abdomen (disorder)                 |
| Fracture subluxation of acromioclavicular joint (disorder)           | Fracture subluxation of joint of upper limb (disorder)   |
| Genital herpes simplex (disorder)                                    | Infectious disease of genitourinary system (disorder)    |
| Plasmodium vivax malaria with rupture of spleen (disorder)           | Infectious disease of abdomen (disorder)                 |
| Open fracture of thoracic spine with spinal cord lesion (disorder)   | Fracture of spine with spinal cord lesion (disorder)     |
| Open injury of diaphragm (disorder)                                  | Open wound of thorax (disorder)                          |
| Osteitis fibrosa cystica generalisata (disorder)                     | Degenerative disorder of bone (disorder)                 |
| Primary malignant neoplasm of fibula (disorder)                      | Malignant neoplasm of long bone of lower leg (disorder)  |

Another cause of false positives is the incorrect existing hierarchical relations in SNOMED CT that our approach leverages to suggest potentially missing hierarchical relations. Table 5.4 shows seven examples of false positives generated by our approach due to the incorrect existing hierarchical relations. For instance, our approach suggests “*Encysted hydrocele of spermatic cord (disorder)*” IS-A “*Soft tissue lesion of pelvic region (disorder)*,” which is incorrect since hydrocele refers to a small “bag of fluid” and is not considered as a soft tissue lesion. This incorrect suggestion is due

**Table 3.4:** Examples of false positives caused by the incorrect existing hierarchical relations.

| False Positives  | Reason: Incorrect Existing Relations   |
|--|--|
| Disorder of left sacroiliac joint (disorder) IS-A Disorder of left lower extremity (disorder)                        | Disorder of pelvic girdle (disorder) IS-A Disorder of lower extremity (disorder)         |
| Encysted hydrocele of spermatic cord (disorder) IS-A Soft tissue lesion of pelvic region (disorder)                  | Encysted hydrocele of spermatic cord (disorder) IS-A Soft tissue lesion (disorder)       |
| Malignant neoplasm of sacral vertebra (disorder) IS-A Malignant neoplasm of bone of lower limb (disorder)            | Neoplasm of sacrum (disorder) IS-A Neoplasm of lower limb (disorder)                     |
| Algodystrophy of foot (disorder) IS-A Degenerative disorder of extremity (disorder)                                  | Algodystrophy (disorder) IS-A Degenerative disorder (disorder)*                          |
| Reflex sympathetic dystrophy of upper extremity (disorder) IS-A Degenerative disorder of extremity (disorder)        | Algodystrophy (disorder) IS-A Degenerative disorder (disorder)*                          |
| Secondary malignant neoplasm of sacrum (disorder) IS-A Secondary malignant neoplasm of bone of lower limb (disorder) | Neoplasm of sacrum (disorder) IS-A Neoplasm of lower limb (disorder)                     |
| Autosomal recessive popliteal pterygium syndrome (disorder) IS-A Dysplasia of limb (disorder)                        | Popliteal pterygium syndrome (disorder) IS-A Congenital anomaly of lower limb (disorder) |

\*: indicates that the incorrect existing relation has been removed in the newer versions of SNOMED CT.

to an existing relation: “*Encysted hydrocele of spermatic cord (disorder)*” IS-A “*Soft tissue lesion (disorder)*.” In addition, there are two false positives caused by the same existing relation: “*Algodystrophy (disorder)*” IS-A “*Degenerative disorder (disorder)*” in the September 2017 release of SNOMED CT US edition that we used. It is worth noting that this relation is no longer existent in the current version of SNOMED CT, that is, this incorrect hierarchical relation has been removed.

### 3.3 Discussion

In this work, we introduce a lexical approach for exhaustive detection of potentially missing hierarchical relations in SNOMED CT. It can be seen that our approach can not only detect intuitive/straightforward relations such as “*Primary adenosquamous cell carcinoma of larynx (disorder)*” IS-A “*Carcinoma of larynx (disorder)*,” but also uncover complicated cases such as “*Genital herpes simplex (disorder)*” IS-A “*Infectious disease of genitourinary system (disorder)*” and “*Plasmodium vivax malaria with rupture of spleen (disorder)*” IS-A “*Infectious disease of abdomen (disorder)*”

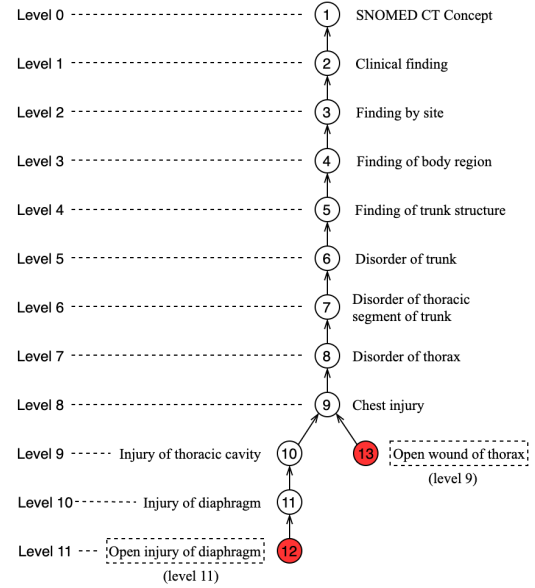
(Table 3.3). Since our approach only requires the hierarchical structure and concept names as the input, it can be generally applied to other terminologies or ontologies.

### 3.3.1 Comparison with previous work

In previous work [26], we introduced a structural-lexical approach for the detection of potentially missing hierarchical relations in SNOMED CT, by leveraging the lexical attributes of concepts in non-lattice subgraphs. A pair of concepts is known as a non-lattice pair if they share more than one maximal common descendant. Non-lattice subgraphs derived from non-lattice pairs often reveal quality issues including missing hierarchical relations. In this work, we perform exhaustive detection of potentially missing hierarchical relations without limiting to the non-lattice substructures.

More importantly, this work identifies previously undiscovered missing hierarchical relations. Among 38,615 potentially missing hierarchical relations identified in this work, 36,534 (94.6%) are newly discovered compared with those in previous work [26]. Among 6,946 potentially missing hierarchical relations from the “*Clinical finding*” sub-hierarchy in this work, 6,081 (87.5%) are newly identified compared with those in previous work [26].

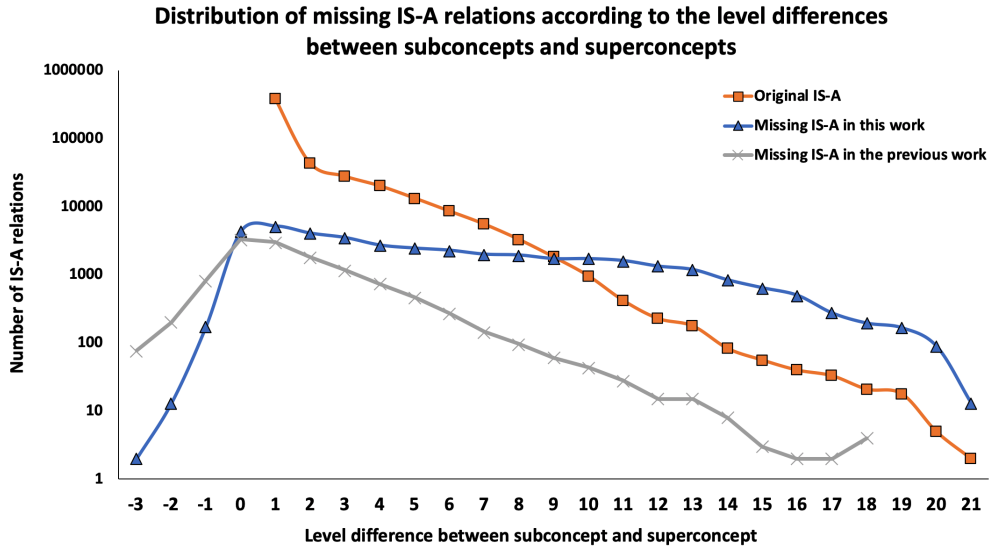
Since this work leverages the entire structure of SNOMED CT while the previous work focuses on non-lattice substructures, it is intuitive to further investigate the level differences of the subconcepts and superconcepts involved in the potentially missing hierarchical relations. Therefore,



**Figure 3.1:** The levels of concepts involved in a missing hierarchical relation: “*Open injury of diaphragm (disorder)*” IS-A “*Open wound of thorax (disorder)*.”

we computed the level of each concept in SNOMED CT (i.e., the number of concepts in the path from the root to the concept). For concepts with multiple paths from the root, we chose the number of the longest path. We considered the root’s level as 0. For instance, Figure 3.1 shows that the level of concept “*Open injury of diaphragm (disorder)*” is 11 and the level of concept “*Open wound of thorax (disorder)*” is 9. The level difference between these two concepts is 2.

We compared the level difference of subconcepts and superconcepts for potentially missing hierarchical relations identified in this work and previous work [26]. Figure 3.2 shows the number of potentially missing hierarchical relations in terms of the level difference between the subconcept and superconcept. The level difference ranges from -3 to 21 in this work and -3 to 18 in previous work. A negative level difference indicates that the superconcept has a higher level than the subconcept does. For the previous work [26], 6% of identified missing hierarchical relations have a level difference that is greater than 5; while for this work, over 32% of the detected missing hierarchical relations have a level difference that is greater than 5. It can also be seen that for



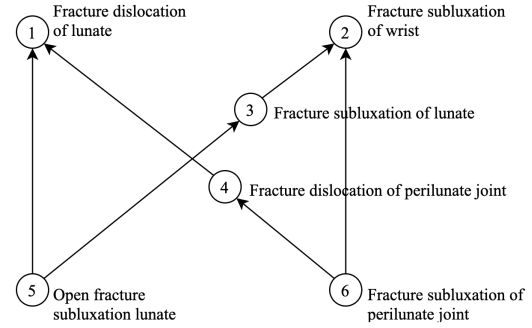
**Figure 3.2:** Distribution of potentially missing hierarchical relations detected in this work and previous work according to the level differences between subconcepts and superconcepts.



each of the non-negative level differences (0 to 21), this work consistently identifies more potentially missing hierarchical relations than the previous work does; while for each of the negative level differences (-3 to -1), the previous work detects more potentially missing hierarchical relations than this work does.

Another major distinction is regarding the construction of lexical features for concepts. In the previous work [26], a concept in a non-lattice subgraph is modeled as a set of words in its FSN with enriched lexical features inherited from its ancestors within the non-lattice subgraph.

For instance, Figure 3.3 shows a non-lattice subgraph identified in the previous work [26], where concept 6, “*Fracture subluxation of perilunate joint*,” has a set of lexical features {*fracture, subluxation, of, perilunate, joint, dislocation, lunate, wrist*}. In this work, we model each concept as a set of words and noun phrases, with enriched lexical features inherited from



**Figure 3.3:** A non-lattice subgraph identified in the previous work [26]. This non-lattice subgraph suggests a missing hierarchical relation between concepts 3 and 1: “*Fracture subluxation of lunate*” IS-A “*Fracture dislocation of lunate*”.

all its ancestors in the entire SNOMED CT. Take the same concept “*Fracture subluxation of perilunate joint*” as an example, this work generates a set of lexical features for the concept as {*fracture, subluxation, of, perilunate, joint, fracture subluxation, perilunate joint, traumatic dislocation, dislocation, traumatic, lunate, lunate bone, bone, wrist, limb structure, finding, structure, limb, upper limb, upper, wrist joint, disorder, fracture dislocation, lesion, musculoskeletal system, injury, system, musculoskeletal, arthropathy, wrist region, region, radiocarpal, radiocarpal joint, body region, body, extremity, upper extremity, traumatic injury, skeletal, skeletal system, connective tissue, tissue, joint injury, body system, bone finding, musculoskeletal finding,*

*joint finding, carpal bone, disease, bone injury*}. As can be seen, concepts have more enriched lexical features to represent their meanings in this work.

### **3.4 Conclusions**

This Chapter presents a lexical-based approach to exhaustively detect potentially missing hierarchical relations in the SNOMED CT. Each concept is modeled as a set of enriched lexical features consisting of words and noun phrases in the name of the concept itself and its ancestors. Pairwise comparison of the concepts' lexical features automatically suggests potentially missing hierarchical relations. The results showed that this approach is effective in identifying missing hierarchical relations. Analysis of false positive cases further revealed incorrect existing hierarchical relations in the SNOMED CT.

## CHAPTER 4. Detecting Missing IS-A Relations in the NCI Thesaurus Using an Enhanced Hybrid Approach

To automatically identify missing hierarchical relations, one commonly used approach is to find features to represent the meanings of concepts [22, 23, 25, 26, 31, 32, 43] and check whether there exist any subsumption relations between the represented meanings. In biomedical ontologies, two important aspects can be utilized to represent the semantic meaning of a concept – lexical features and logical definitions.

Lexical features have been widely adopted to detect missing hierarchical relations in ontologies including the NCI Thesaurus, Gene Ontology and SNOMED CT (e.g., auditing method introduced in Chapter 3). However, in many cases, it is challenging to get the machine to catch the meanings and other details behind the words. Take concept “*Sarcoma*” in the NCI Thesaurus as an example. Purely from the concept name itself, the machine will not be able to know that this concept refers to a malignant neoplasm of the soft tissue or bone. In addition, concept names are defined manually by curators of biomedical ontologies and inconsistencies may exist during the naming process [42], which may further affect the subsumption checking.

When it comes to logical definitions (or role definitions), they are formally defined and often contain meanings beyond concept names. Consider the previous example “*Sarcoma*.” Table 4.1 shows its role definitions in the NCI Thesaurus, including a subtype relation (*IS-A*, *Malignant Neoplasm*) and an associative role (i.e., attribute relation) (*Disease\_Has\_Associated\_Anatomic\_Site*, *Connective and Soft Tissue*). However, logical definitions (or role definitions) are often incomplete, making them impractical to be solely used in representing meanings of concepts. For instance, in the 19.08d version of the NCI Thesaurus, only 17,052 out of 146,688 (11.62%) concepts are considered as fully defined in logical definition; and in the 11/02/2019 release of Gene Ontology, the number is 12,011 out of 44,650 (26.9%).

To derive more precise missing hierarchical relations from subsumption testing, this Chapter presents a hybrid semantic model that leverages both lexical features and role definitions, aiming at providing more comprehensive information while representing the meanings of concepts.

**Table 4.1:** The role definitions of concept “*Sarcoma*” (*C9118*) in the NCI Thesaurus [3].

| Attribute type                        | Value                                   |
|---------------------------------------|---|
| IS-A                                  | Connective and Soft Tissue Neoplasm     |
| IS-A                                  | Malignant Neoplasm                      |
| Disease_Has_Abnormal_Cell             | Malignant Cell                          |
| Disease_Has_Abnormal_Cell             | Neoplastic Cell                         |
| Disease_Excludes_Normal_Cell-Origin   | Epithelial Cell                         |
| Disease_Excludes_Normal_Tissue-Origin | Epithelial Tissue                       |
| Disease_Has_Associated_Anatomic_Site  | Connective and Soft Tissue              |
| Disease_Has_Normal_Tissue-Origin      | Connective and Soft Tissue              |
| Disease_Excludes_Finding              | Benign Cellular Infiltrate              |
| Disease_Excludes_Finding              | Indolent Clinical Course                |
| Disease_Excludes_Finding              | Intermediate Filaments Present          |
| Disease_Excludes_Finding              | Intracytoplasmic Eosinophilic Inclusion |
| Disease_Has_Finding                   | Malignant Cellular Infiltrate           |

## 4.1 Methods

In this work, we focus on detecting missing hierarchical relations for non-lattice subgraphs which often contain quality issues [10, 22, 23, 26]. To identify missing hierarchical relations in non-lattice subgraphs, we first find a proper way to represent the meanings of concepts, and then check whether there exists any subsumption relations between the represented meanings of unlinked concepts (i.e., not connected by hierarchical relations either directly or transitively) within non-lattice subgraphs.

There are mainly three steps: (1) compute non-lattice subgraphs and identify candidate pairs of concepts that are currently not linked by hierarchical relations;

(2) for each concept, construct a model that harmonizes associative roles, words and roots of noun chunks within its concept name and its ancestor’s names, to represent its meaning; (3) perform subsumption checking for candidate pairs based on our hybrid model.

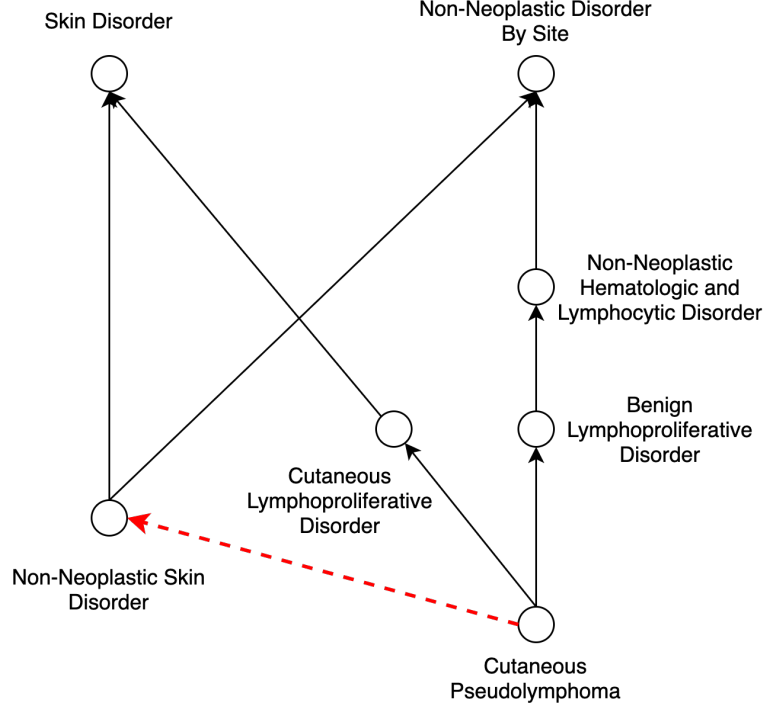
#### 4.1.1 Computing non-lattice subgraphs and generating candidate pairs

Concepts in an ontology are organized into an IS-A hierarchy, which can be considered as a directed acyclic graph. Given two concepts  $A$  and  $B$  in the ontology, a common ancestor  $X$  of  $A$  and  $B$  is known as their *minimal* common ancestor, if  $A$  and  $B$  do not have any other common ancestor  $Y$  such that  $X$  is an ancestor of  $Y$ . Similarly, a common descendant  $P$  of  $A$  and  $B$  is known as their *maximal* common descendant, if  $A$  and  $B$  do not have any other common descendant  $Q$  such that  $P$  is a descendant of  $Q$ . An ontology forms a *lattice* if any two concepts in the ontology have a unique minimal common ancestor and a unique maximal common descendant. Lattice is a desirable property for a well-formed ontology or terminology [64].

A pair of concepts is called a *non-lattice pair* if the two concepts have more than one maximal common descendant. A *non-lattices subgraph* can be obtained from a non-lattice pair by first reversely computing the minimal common ancestors of the maximal common descendants of the non-lattice pair; and then aggregating all the concepts and hierarchical relations between them [22]. Figure 4.1 shows a non-lattice subgraph in the NCI Thesaurus (19.08d version) obtained from the non-lattice pair (“*Skin Disorder*,” “*Non-Neoplastic Disorder By Site*”) with two maximal common descendants “*Non-Neoplastic Skin Disorder*” and “*Cutaneous Pseudolymphoma*”.

We leverage an efficient non-lattice extraction algorithm [100] to compute all the non-lattice subgraphs in the NCI Thesaurus. Then we identify potentially missing hierarchical relations between pairs of concepts (denoted as candidate pairs) which are currently not linked by hierarchical relations in the non-lattice subgraphs. Take

the non-lattice subgraph shown in Figure 4.1 as an example, (“*Non-Neoplastic Skin Disorder*,” “*Cutaneous Pseudolymphoma*”) is a candidate pair and (“*Skin Disorder*,” “*Benign Lymphoproliferative Disorder*”) is another.



**Figure 4.1:** An example of non-lattice subgraphs in the 19.08d version of NCI Thesaurus. Concepts are connected by hierarchical relations. The red dotted line shows a potentially missing hierarchical relation between concepts “*Cutaneous Pseudolymphoma*” and “*Non-Neoplastic Skin Disorder*” identified by our method.

#### 4.1.2 Modeling concepts

In this work, we introduce a comprehensive semantic model that utilizes role definitions and lexical features to represent the meanings of concepts. Given a concept  $C$ , its semantic model contains five parts ( $C_{bow}$ ,  $C_{ebow}$ ,  $C_r$ ,  $C_{er}$ ,  $C_a$ ):

1. bag-of-words  $C_{bow}$ , which includes words appearing in its preferred name;
2. enriched bag-of-words  $C_{ebow}$ , which includes words appearing in its preferred name and words in its ancestors’ preferred names;

3. roots of noun chunks  $C_r$ , which includes roots of noun chunks in its preferred name;
4. enriched roots of noun chunks  $C_{er}$ , which includes roots of noun chunks in its preferred name and in its ancestors' preferred names; and
5. associative roles  $C_a$ .

Figure 4.2 shows the semantic models for concepts “*Cutaneous Pseudolymphoma*” and “*Non-Neoplastic Skin Disorder*” in the non-lattice subgraph shown in Figure 4.1.

| Cutaneous Pseudolymphoma (C62776)     |  |                      |                               |  |
|---------------------------------------|--|----------------------|-------------------------------|--|
| Bag-of-Words                          | Enriched Bag-of-Words  | Roots of Noun Chunks | Enriched Roots of Noun Chunks | Associative Roles  |
| cutaneous, pseudolymphoma             | cutaneous, pseudolymphoma, benign, lymphoproliferative, disorder, skin, non-neoplastic | pseudolymphoma       | pseudolymphoma, disorder      | (Disease_Has_Primary_Anatomic_Site, Skin),<br>(Disease_Has_Finding, Perivascular Lymphocytic Infiltrate),<br>(Disease_May_Have_Finding, Regression),<br>(Disease_May_Have_Finding, Papular Lesion),<br>(Disease_Has_Associated_Anatomic_Site, Skin),<br>(Disease_Has_Associated_Anatomic_Site, Integumentary System),<br>(Disease_Has_Associated_Anatomic_Site, Hematopoietic and Lymphatic System),<br>(Disease_Has_Primary_Anatomic_Site, Lymphatic System),<br>(Disease_Has_Primary_Anatomic_Site, Hematopoietic and Lymphatic System),<br>(Disease_Has_Normal_Tissue_Origin, Lymphoid Tissue),<br>(Disease_Has_Normal_Tissue_Origin, Hematopoietic and Lymphoid Tissue),<br>(Disease_Has_Normal_Cell_Origin, Lymphocyte),<br>(Disease_Has_Normal_Cell_Origin, Hematopoietic and Lymphoid Cell),<br>(Disease_Has_Abnormal_Cell, Abnormal Lymphocyte),<br>(Disease_Has_Finding, Non-Malignant Cellular Infiltrate),<br>(Disease_Has_Finding, Cutaneous Involvement),<br>(Disease_Excludes_Normal_Cell_Origin, Myeloid Cell),<br>(Disease_Excludes_Molecular_Abnormality, Clonal Antigen Receptor Gene Rearrangement) |
| Non-Neoplastic Skin Disorder (C27555) |  |                      |                               |  |
| Bag-of-Words                          | Enriched Bag-of-Words  | Roots of Noun Chunks | Enriched Roots of Noun Chunks | Associative Roles  |
| non-neoplastic, skin, disorder        | non-neoplastic, skin, disorder   | disorder             | disorder                      | (Disease_Has_Associated_Anatomic_Site, Integumentary System),<br>(Disease_Has_Primary_Anatomic_Site, Skin),<br>(Disease_Has_Finding, Cutaneous Involvement)  |

**Figure 4.2:** Semantic models of concepts “*Cutaneous Pseudolymphoma (C62776)*” and “*Non-Neoplastic Skin Disorder (C27555)*” which are contained in the non-lattice subgraph shown in Figure 4.1.

Note that subtype relations in the role definitions are not included in the semantic model, since our goal is to identify missing hierarchical relations (i.e., subtype relations). Alternatively, we use features inherited from concept’s ancestor (i.e.,  $C_{ebow}$  and  $C_{er}$ ) to embody the subtype relations, which could gather more concept-related information and thus help refine the meanings of concepts. We maintain both original lexical features (i.e.,  $C_{bow}$  and  $C_r$ ) and enriched ones (i.e.,  $C_{ebow}$  and  $C_{er}$ ) for performing subsumption testing later.

#### 4.1.2.1 Lexical features

The regular bag-of-words  $C_{bow}$  and enriched bag-of-words  $C_{ebow}$  could convey the meaning of a concept to some extent. However, there exist some words that may express different meanings depending on the contexts under which they appear. For example, word “erlotinib” in concept “*Erlotinib*” and in concept “*Erlotinib Hydrochloride*” convey different meanings – the former refers to the chemical item itself while the latter is used to describe word “hydrochloride.” Therefore, even though both concepts contain the same word “erlotinib,” it should be considered as a different lexical feature for each concept.

To handle such cases (i.e., a noun used as a descriptive term), our idea is to leverage a technique in Natural Language Processing (NLP) called dependency parsing which could extract the grammatical structure and relationships between words for a given phrase. For example, after parsing concept name “*Malignant Bladder Neoplasm*,” we can get “malignant bladder neoplasm” whole as a noun chunk. The word “malignant” is used to modify “neoplasm” in terms of the type (i.e., benign or malignant) while the word “bladder” is used to modify “neoplasm” in term of the location (i.e., anatomic site). In this work, besides bag-of-words  $C_{bow}$  (and enriched  $C_{ebow}$ ), we also adopt roots of noun chunks  $C_r$  (and enriched  $C_{er}$ ) as part of the lexical feature. Given a concept name, we use spaCy [101], an open-source library for NLP, to parse it and recognize the roots of noun chunks. In the previous example, “neoplasm” is denoted as a root of noun chunk since other words are used to modify it. By utilizing roots of noun chunks  $C_r$ , to some extent we could distinguish different meanings of a word in different contexts. In the concepts “*Erlotinib*” and “*Erlotinib Hydrochloride*,” word “erlotinib” will be taken as two different words – a root of noun chunk in the former concept, but a descriptive term (i.e., not a root of noun chunk) in the latter concept.

In this work, we also adopt a list of stop words that may distort the represented meanings of concepts. As mentioned in our previous work [39], concept names which



contain “and” are often inconsistent with what they actually mean and their role definitions. For example, concept “*Lip **and** Oral Cavity Squamous Cell Carcinoma*” actually refers to a squamous cell carcinoma arising from the lip **or** the oral cavity. In this work, we do not perform subsumption testing for candidate pairs that include concepts whose  $C_{bow}$  contain any stop word. In addition, while generating enriched lexical features  $C_{ebow}$  and  $C_{er}$ , concepts will not inherit lexical features  $C_{bow}$  and  $C_r$  from their ancestors containing any stop word such that the stop words will not propagate. More specifically, as long as an ancestor contains a stop word, none of the ancestor’s lexical features will be inherited. The list of stop words used in this step is the same as the one introduced in Chapter 3.

In Figure 4.2, it can be seen that concept “*Cutaneous Pseudolymphoma*” has two single words and inherits seven words from its ancestors such as “*Benign Lymphoproliferative Disorder*,” “*Skin Disorder*” and “*Non-Neoplastic Disorder*,” which enrich the meaning expressed by the concept name. Also, word “pseudolymphoma” is recognized as the root of noun chunk “cutaneous pseudolymphoma.” Concept “*Cutaneous Pseudolymphoma*” also inherits another root of noun chunk “disorder” from its ancestor “*Benign Lymphoproliferative Disorder*.” Note that another ancestor of concept “*Cutaneous Pseudolymphoma*” is “*Non-Neoplastic Hematologic and Lymphocytic Disorder*,” which contains a stop word “and.” Hence,  $C_{bow}$  and  $C_r$  of this ancestor are not inherited.

#### 4.1.2.2 Associative roles

In our model, we use associative roles  $C_a$  to collect and adjust the meaning of concepts that may not be fully expressed by lexical features, especially for concepts that are lexically similar but should not be linked by hierarchical relations. Consider the concepts “*Metastatic Malignant Neoplasm in the Pancreas*” and “*Metastatic Malignant Pancreatic Neoplasm*.” If only lexical features are considered, the former

concept’s lexical features include all of the latter one’s after the enrichment (e.g., “*Metastatic Malignant Neoplasm in the Pancreas*” inherits “pancreatic” from its ancestor “*Pancreatic Neoplasm*”). However, the former concept “*Metastatic Malignant Neoplasm in the Pancreas*” refers to a malignant neoplasm that has spread to the pancreas from another anatomic site, while “*metastatic malignant pancreatic neoplasm*” actually refers to a malignant neoplasm that arises from the pancreas and has metastasized to another anatomic site. Thus, there should not be any subsumption relations between these two concepts. However, the difference between the two concepts can not be caught purely from their lexical features. To compensate for this, we adopt associative roles which usually contain information that is not included in the literal meanings. Consider the previous example, the former concept has associative role (*Disease\_Has\_Metastatic\_Anatomic\_Site*, *Pancreas*), but the latter concept has role definition (*Disease\_Excludes\_Metastatic\_Anatomic\_Site*, *Pancreas*). Depending on the inclusion and exclusion of metastatic anatomic locations provided by the role definitions could easily distinguish these two concepts.

In this work, to gather as much information as possible, the associative roles we adopted for a concept are the inferred ones that include associative roles inherited from the concept’s ancestors. For instance, in Figure 4.2, concept “*Cutaneous Pseudolymphoma*” contains 18 associative roles (14 inherited from its ancestors), while concept “*Non-Neoplastic Skin Disorder*” contains three associative roles (two inherited from its ancestors).

### 4.1.3 Identifying potentially missing hierarchical relations

As mentioned earlier, in this work, our task is to identify potentially missing hierarchical relations among candidate pairs – pairs of concepts that are not linked by hierarchical relations within non-lattice subgraphs. For each candidate pair ( $A$ ,  $B$ ), we perform a two-step subsumption checking to see if the meaning represented by the

hybrid model of  $A$  is more detailed than  $B$ 's (i.e.,  $A$  IS-A  $B$ ), or vice versa (i.e.,  $B$  IS-A  $A$ ).

In the first step, we perform a lexical-feature-based checking. We consider original lexical features (i.e.,  $C_{bow}$ ,  $C_r$ ) as minimal satisfying features for a concept. In other words, if  $A$ 's enriched lexical features (i.e., all meanings from lexical features that hold for  $A$ ) satisfy  $B$ 's original lexical features, we consider  $A$  is more detailed than  $B$  in terms of lexical features. Here, we do not consider enriched lexical features of  $B$  because  $A$  can then also inherit lexical features from  $B$ 's ancestors if  $A$  becomes a subtype of  $B$ . As we represent lexical features of concepts as sets of words, we simply use set inclusion testing, that is, if  $A$ 's enriched bag-of-words (i.e.,  $A_{ebow}$ ) is a superset of  $B$ 's bag-of-words (i.e.,  $B_{bow}$ ) and  $A$ 's enriched roots of noun chunks (i.e.,  $A_{er}$ ) is a superset of  $B$ 's roots of noun chunks (i.e.,  $B_r$ ), then  $A$  is considered more detailed than  $B$  in lexical feature wise.

In the second step, we perform a role-based checking. To do so, we require that each of the two concepts within a candidate pair should contain at least one associative role and associative roles of two concepts should not be totally identical (otherwise we can not decide which one is more detailed). Further, we check that for each associative role ( $role_B$ ,  $value_B$ ) of  $B$ , if there exists a corresponding role ( $role_A$ ,  $value_A$ ) of  $A$  such that  $role_A$  and  $role_B$  are the same and  $value_B$  is the same or more general than  $value_A$  (i.e.,  $value_B$  is an ancestor of  $value_A$ ). If this is the case, then  $A$  is considered more detailed than  $B$  in terms of role definitions.

If  $A$  is more detailed than  $B$  in terms of both lexical features and role definitions, we consider “ $A$  IS-A  $B$ ” as a potentially missing hierarchical relation. For example, consider a candidate pair (“*Cutaneous Pseudolymphoma*,” “*Non-Neoplastic Skin Disorder*”) in Figure 4.2. “*Cutaneous Pseudolymphoma*” is more detailed than “*Non-Neoplastic Skin Disorder*” in terms of lexical features because the enriched bag-of-words of “*Cutaneous Pseudolymphoma*,” {cutaneous, pseudolymphoma, be-

nign, lymphoproliferative, disorder, skin, non-neoplastic}, is a superset of bag-of-words of “*Non-Neoplastic Skin Disorder*,” {non-neoplastic, skin, disorder}; and the enriched roots of noun chunks of “*Cutaneous Pseudolymphoma*,” {pseudolymphoma, disorder}, is also a superset of roots of noun chunks of “*Non-Neoplastic Skin Disorder*,” {disorder}. In addition, “*Cutaneous Pseudolymphoma*” is more detailed than “*Non-Neoplastic Skin Disorder*” in role definitions, since for each associative role of “*Non-Neoplastic Skin Disorder*,” there is a corresponding role of “*Cutaneous Pseudolymphoma*” that is equivalent or more detailed. Therefore, our approach suggests “*Cutaneous Pseudolymphoma IS-A Non-Neoplastic Skin Disorder*” as a potentially missing hierarchical relation. Note that this missing hierarchical relation has been confirmed by experts from NCI Enterprise Vocabulary Service (EVS) and included in the newer versions of the NCI Thesaurus.

In some cases, a potentially missing hierarchical relation detected could actually be a relation similar to a hierarchical relation, such as “*part of*.” NCI Thesaurus provides associations (i.e., different things from role definitions) between concepts, such as “*Has\_Salt\_Form*,” “*Has\_Target*,” “*Has\_Pharmaceutical\_Transformation*,” etc. We further utilize them to distinguish those like-hierarchical relations. Given a potentially missing hierarchical relation identified by our approach, if two concepts are already linked by any kind of these associations, then the missing hierarchical relation will be abandoned.

Another thing to consider is that due to the large size of some non-lattice subgraphs, there may exist an overlap between non-lattice subgraphs which may result in redundant missing hierarchical relations being suggested. We adopt the same strategy in Chapter 3 to remove redundant relations which can be inferred by other missing or existing hierarchical relations.

## 4.2 Results

We applied our enhanced hybrid approach to the 19.08d inferred version of the NCI Thesaurus for identifying potentially missing hierarchical relations.

### 4.2.1 Non-lattice subgraphs and suggested hierarchical relations

In total, 10,216 non-lattice subgraphs were obtained in 16 sub-hierarchies of the NCI Thesaurus. 55 non-redundant missing hierarchical relations were suggested for five sub-hierarchies. Table 4.2 shows the number of suggested missing hierarchical relations for each of the five sub-hierarchies. For example, 34 non-redundant missing hierarchical relations were suggested in the “*Disease, Disorder or Finding*” sub-hierarchy.

**Table 4.2:** The number of potentially missing hierarchical relations identified for sub-hierarchies.

| Sub-hierarchy                               | # of Non-lattice Subgraphs | # of Suggested Missing IS-A relations |
|---|----------------------------|---------------------------------------|
| Disease, Disorder or Finding                | 8,075                      | 34                                    |
| Experimental Organism Diagnosis             | 257                        | 18                                    |
| Drug, Food, Chemical or Biomedical Material | 922                        | 1                                     |
| Molecular Abnormality                       | 143                        | 1                                     |
| Activity                                    | 109                        | 1                                     |

### 4.2.2 Evaluation

For evaluation, we provided the NCI EVS domain experts, who manage the NCI Thesaurus, with 55 potentially missing hierarchical relations identified by our approach. 29 out of 55 were confirmed by EVS experts and have been incorporated in the newer version of the NCI Thesaurus. Table 4.3 lists ten examples of valid missing hierarchical relations verified by EVS experts, including “*Glycine Encephalopathy*” IS-A “*Congenital Nervous System Disorder*” and “*Congenital Vena Cava Abnormality*” IS-A “*Congenital Cardiovascular Abnormality*.”

**Table 4.3:** Ten examples of valid missing hierarchical relations confirmed by EVS experts.

| Subconcept  | Superconcept                                |
|---|---|
| Glycine Encephalopathy                                      | Congenital Nervous System Disorder          |
| Tumor Infiltrating Lymphocytes-N2-Transduced                | Therapeutic Tumor Infiltrating Lymphocytes  |
| Stage 0 Anal Cancer AJCC v8                                 | Anal Precancerous Condition                 |
| Cutaneous Pseudolymphoma                                    | Non-Neoplastic Skin Disorder                |
| Congenital Vena Cava Abnormality                            | Congenital Cardiovascular Abnormality       |
| Mouse Cardiac Fibrosarcoma                                  | Mouse Cardiac Sarcoma                       |
| Fibrosarcoma of the Mouse Intestinal Tract                  | Mouse Malignant Mesenchymal Neoplasm        |
| Carcinoma of the Mouse Larynx                               | Mouse Carcinoma                             |
| Eyelid Xanthoma   | Non-Neoplastic Eyelid Disorder              |
| Autoimmune Lymphoproliferative Syndrome-Associated Lymphoma | Immunodeficiency-Related Malignant Neoplasm |

### 4.3 Discussion

In this work, we combine role definitions and lexical features to suggest missing hierarchical relations in the NCI Thesaurus. The evaluation results show that our hybrid approach is promising in identifying missing hierarchical relations. From the true positives, such as “*Glycine Encephalopathy*” IS-A “*Congenital Nervous System Disorder*” and “*Cutaneous Pseudolymphoma* IS-A *Non-Neoplastic Skin Disorder*,” we find that using enriched lexical features for subconcepts help recognize meanings related to the concepts that cannot be caught from their own concept names.

#### 4.3.1 Analysis of false positives

Even though this approach correctly suggested missing hierarchical relations in majority of the cases (i.e., 29 out of 55), there were still cases where the approach made incorrect suggestions. By reviewing such invalid suggestions, we identified two major causes for them.

The first cause is that the existence of erroneous hierarchical relations in NCI Thesaurus has led to invalid missing hierarchical suggestions. For example, our approach suggested “*Carcinosarcoma of the Mouse Prostate Gland*” IS-A “*Carcinoma*

of the *Mouse Prostate Gland*” mainly based on an existing hierarchical relation “*Carcinoma of the Mouse Prostate Gland*” IS-A “*Mouse Carcinoma*.” However, as stated by EVS experts, “carcinoma” is not a kind of “carcinoma.” Thus, the existing hierarchical relation on which we rely to derive the missing hierarchical relation is incorrect, and it has been fixed by EVS experts in the newer release of the NCI Thesaurus. In total, 7 out of 26 false positive cases fall into this cause. In such cases, even though our suggestions of missing hierarchical relations were incorrect, they further revealed problems within the existing hierarchy of the NCI thesaurus that in turn help improve the quality of the NCI thesaurus.

Secondly, since we only adopted original lexical features ( $C_{bow}$ ,  $C_r$ ) for superconcepts during subsumption testing, the meanings beyond the original lexical features and logical definitions could lead incorrect missing hierarchical relations to be suggested. Consider the false positive “*Diffuse Pulmonary Lymphangiomatosis*” IS-A “*Pulmonary Vascular Disorder*.” The subconcept is a kind of “neoplasm,” however, the superconcept has an ancestor “*Non-Neoplastic Lung Disorder*.” Since a neoplasm could not be a subtype of a non-neoplastic disorder, this suggestion is invalid. Other similar cases include: “*Conjunctival Kaposi Sarcoma*” IS-A “*Conjunctival Vascular Disorder*” and “*Retinal Hemangioma*” IS-A “*Retinal Vascular Disorder*.” Since meanings like “non-neoplastic” could be found in the enriched lexical features of superconcepts (i.e., inherited from ancestors), a natural question would be: Whether adopting enriched lexical features for both concepts within candidate pairs during lexical-based subsumption testing could improve the performance of our method?

To study this, we further utilized enriched lexical features of superconcept and subconcepts in lexical-based subsumption checking. Therefore, in order for a hierarchical relation to be suggested, the enriched lexical features of the subconcept now should also contain the original lexical features of the superconcept’s ancestors. In total, 45 missing hierarchical relations were identified in this setting. The result

was found to be a subset of our previous result. One exception is that a missing hierarchical relation was considered redundant but became non-redundant as some missing hierarchical relations are no longer included in the result. Since the hierarchical relation was redundant to a valid hierarchical relation, this hierarchical relation is also considered as valid. Among those 45 missing hierarchical relations, 29 were valid hierarchical relations, the number of true positives went down by 3 but the number of false positives went down by 7. We noticed that some false positives in the format of “neoplasm” IS-A “non-neoplastic” still appeared in the result because the role definitions of the superconcepts are not sufficient (i.e., incompleteness). For example, “*Kidney Lymphangioma*” IS-A “*Kidney Vascular Disorder*.” The superconcept “*Kidney Vascular Disorder*” should be a “non-neoplastic” disorder, however, in the role definitions, none of its ancestors is “non-neoplastic” disorder and none of its associative roles indicates that it is not a kind of “neoplasm.” Another example is “*Brain Astrocytoma*” IS-A “*Brain Disorder*.” Therefore, adopting enriched lexical features for both superconcept and subconcept during lexical-based subsumption checking could improve the performance, but only slightly due to the incompleteness of role definitions.

### 4.3.2 Comparison with other approaches

The hybrid approach introduced in this Chapter can be categorized as a structural-lexical-based method introduced in Chapter 2.

Compared with structural-based approaches such as abstraction network (AbNs) [33, 36] which often require domain expert to manually review problematic areas in ontologies to reveal the exact quality issues, our approach not only identifies problematic areas (i.e., non-lattice subgraphs), but also automatically suggests missing hierarchical relations (the actual quality issues) in the problematic areas.

In previous lexical-based and structural-lexical-based methods, bag-of-words model



is often adopted while representing concepts [25, 26, 43, 102]. However, individual words may not be sufficient to represent the semantic meaning of a concept. In contrast, our hybrid model leverages both lexical features and associative roles to provide more comprehensive information for concepts’ meanings.

### 4.3.3 Comparison with our previous work

In our previous work [102], we developed a lexical-based approach to identify missing hierarchical relations in the NCI Thesaurus. The lexical feature used in that work was the enriched bag-of-words (i.e.,  $C_{ebow}$  in this work). Since only one kind of lexical features was used, several other restrictions such as stop words, antonym pairs and location restrictions were applied to avoid potential false identification of missing hierarchical relations. In total, 925 potentially missing hierarchical relations were identified from 9,512 non-lattice subgraphs in 19.01d inferred version of the NCI Thesaurus. We provided EVS experts with 253 potentially missing hierarchical relations in non-lattice subgraphs of size less than or equal to 15. EVS experts confirmed 73 out of 253 suggested missing hierarchical relations. We compared our hybrid approach in this work with the lexical-based approach in previous work [102] in two aspects.

First, we applied our hybrid approach in this work to the 19.01d inferred version of the NCI Thesaurus. In total, 87 non-redundant missing hierarchical relations were identified, 56 out of which were obtained from non-lattice subgraphs of size less than or equal to 15. Compared with previously evaluated 253 missing hierarchical relations, 28 out of 55 were overlapping. Among those 28 overlapped ones, 14 of them were true positives. Based on this, the precision was improved while the recall was lowered.

In our previous work [102], only one type of lexical features (enriched bag-of-words) was used. Therefore, in the second experiment, we tried to consider associative roles and roots of noun chunks as additional subsumption testing (i.e., in addition to lex-

ical features and other restrictions used in [102]) to further check their effectiveness in helping identify missing hierarchical relations. Given 253 missing hierarchical relations identified in our previous work, 135 out of 253 were the cases in which both the subconcept and the superconcept contained at least one associative roles and their associative roles were not identical. 32 out of 253 cases satisfied the role-based testing, and 14 out of them were valid ones. When it comes to using roots of noun chunks, 245 cases passed the testing, where 70 out of them were valid ones. When it comes to performing additional subsumption testing based on both features, 31 out of 253 passed the testing, where 14 out of them were valid ones. The results indicate that associative roles can be used as the main tester to recognize differences in the intended meanings of concepts and roots of noun chunks can be used to catch the subtle differences.

#### **4.4 Conclusions**

In this chapter, we introduced a hybrid semantic model that combines lexical features and role definitions of concepts to identify missing hierarchical relations within non-lattice subgraphs in the NCI Thesaurus. The results showed that our approach is capable of uncovering valid missing hierarchical relations. Further examination of false positives revealed erroneous existing hierarchical relations as well as incomplete concept definitions, which in turn also helped improve the quality of the NCI thesaurus. Comparison with our previous lexical-based work further showed the usefulness of leveraging role definitions.

## **CHAPTER 5. A Transformation-based Method for Auditing the IS-A Hierarchy of Biomedical Terminologies in the Unified Medical Language System**

The Unified Medical Language System (UMLS) integrates various source ontologies to support interoperability between biomedical information systems. This chapter presents a novel transformation-based auditing method that leverages the knowledge in the UMLS to systematically identify missing hierarchical relations in its source ontologies. Unlike the traditional ontology auditing methods that often rely on internal knowledge (e.g., approaches introduced in Chapter 3 and Chapter 4), this method leverages not only the ontology itself but also the knowledge from other multiple ontologies in the UMLS (i.e., both internal and external knowledge). This will result in newly identified missing hierarchical relations that would not be uncovered by only looking into one or two individual ontologies.

### **5.1 Methods**

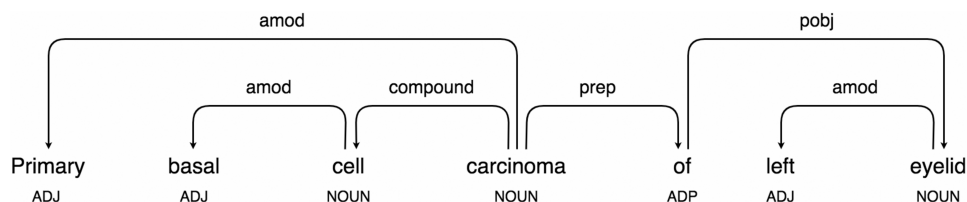
This work is based on the UMLS 2019AB release. A large proportion of concept names (or atoms) in the UMLS contain more than one noun chunk. The key idea of this transformation-based auditing method is to replace those noun chunks in a concept name with more general terms. If a newly generated name after the replacement is an existing concept name in the same source ontology, then we consider there is a potentially missing hierarchical relation between the two concepts corresponding to the original and new concept names.

This method consists of four main steps to identify potentially missing hierarchical relations for each concept name in the UMLS: (1) parse the concept name and identify noun chunks; (2) generate replacement candidates for noun chunks; (3) perform concept name transformation and construct new potential concept names; and

(4) map newly constructed concept names to atoms and identify potentially missing hierarchical relations in the source ontologies.

### 5.1.1 Parsing concept names

We first convert each concept name to lower case. We then use spaCy [101], an open-source library for advanced NLP, to perform dependency parsing and identify noun chunks within concept names. For example, Figure 5.1 shows the dependency graph of the concept name “*Primary basal cell carcinoma of left eyelid*” where two base noun chunks can be identified: “primary basal cell carcinoma” and “left eyelid.” Here a base noun chunk consists of a head (e.g., “carcinoma”) plus words describing the head (e.g., “primary basal cell”) [103]. Note that “basal cell” is not a base noun chunk since it is used to modify or describe “carcinoma.” Instead, we consider such noun phrases describing the head as secondary noun chunks.



**Figure 5.1:** Dependency graph of the concept name “*Primary basal cell carcinoma of left eyelid*.”

After the parsing, each concept name  $C$  can be represented as an ordered array of elements  $[c_1, c_2, \dots, c_n]$ , where  $c_i$  can be a single word, a base noun chunk, or a secondary noun chunk. For instance, the concept name “*Primary basal cell carcinoma of left eyelid*” can be represented in two forms: (1) [primary basal cell carcinoma, of, left eyelid]; and (2) [primary, basal cell, carcinoma, of, left eyelid].

### 5.1.2 Identifying replacement candidates

In this step, we identify replacement candidates that are more general than the noun chunks (base and secondary) in each concept name. If a noun chunk can be mapped to a UMLS atom (i.e., the noun chunk is also a concept name in an existing source ontology), then we consider the concept names of this atom’s ancestors in its source ontology as replacement candidates for the noun chunk; otherwise, the noun chunk is considered as not having any replacement candidates. In other words, we leverage existing hierarchical relations in the UMLS source ontologies to identify replacement candidates. To avoid replacement candidates being too general, we leveraged ancestors of the atom within a distance of two levels using Depth-limited-search[104].

Take the concept name “*Acute dacryoadenitis of left eye*” in Table 5.1 as an example, it can be represented as an array [acute dacryoadenitis, of, left eye]. The noun chunk “acute dacryoadenitis” can be mapped to 9 atoms. For example, *A2889158* is an atom sourced from the SNOMED CT (US edition) with seven level-2 ancestors. After going through all the 9 atoms, the following replacement candidates for “acute dacryoadenitis” can be obtained: “*disorder of lacrimal gland,*” “*disorder of eyelid or lacrimal system,*” “*dacryoadenitis,*” “*inflammation of specific body systems,*” “*acute inflammatory disease,*” “*inflammatory disorder of head,*” “*acute disease,*” and “*inflammatory disorder.*”

**Table 5.1:** An example of the transformation process for concept name “*Acute dacryoadenitis of left eye.*”

|   |  |
|---|--|
| <b>Concept name</b>   | Acute dacryoadenitis of left eye   |
| <b>Representation</b> ( $[c_1, c_2, c_3]$ )                                 | [acute dacryoadenitis, of, left eye]   |
| <b>Replacement candidates for “acute dacryoadenitis” (<math>r_1</math>)</b> | {dacryoadenitis, inflammation of specific body systems, acute disease, acute inflammatory disease, inflammatory disorder, inflammatory disorder of head, disorder of eyelid or lacrimal system, disorder of lacrimal gland}  |
| <b>Replacement candidates for “left eye” (<math>r_3</math>)</b>             | {organ of special sense, eye, subdivision of face}   |
| <b>Combinatorial replacements</b>   | [[{acute dacryoadenitis, dacryoadenitis, inflammation of specific body systems, acute disease, acute inflammatory disease, inflammatory disorder, inflammatory disorder of head, disorder of eyelid or lacrimal system, disorder of lacrimal gland}, of, {left eye, organ of special sense, eye, subdivision of face}] |
| <b>Potentially missing IS-A relations detected in source ontologies</b>     | SNOMEDCT.US:<br>“acute dacryoadenitis of left eye” IS-A “acute disease of eye”<br>MEDCIN:<br>“acute dacryoadenitis of left eye” IS-A “inflammatory disorder of eye”  |

### 5.1.3 Concept name transformation

For each concept name with noun chunk(s) such that the replacement candidates have been identified already, we replace the original noun chunk(s) with their corresponding candidates to generate new potential concept names, which may serve as supertypes of the original concept name (since the replacement candidates are more general than the original noun chunk). Formally, given a concept name  $C$  represented by  $[c_1, c_2, \dots, c_n]$  where there exists an  $i$  such that  $c_i$  is a base or secondary noun chunk and  $r_i$  is a set of replacement candidates for  $c_i$ , then we replace  $c_i$  with any candidate in  $r_i$  and concatenate the array as a string to construct new concept names that may serve as  $C$ ’s supertypes. If there are multiple such  $i$ ’s, we will perform combinatorial replacements for multiple  $i$ ’s.

Take the concept name “*Acute dacryoadenitis of left eye*” in Table 5.1 as an example. There are three elements in its array representation  $[c_1, c_2, c_3]$  where  $c_1$  and  $c_3$  are base noun chunks. There are 8 replacement candidates for  $c_1$  and 3 for  $c_3$ . A total of 35 new potential concept names can be obtained after the combinatorial

replacements for  $c_1$  and  $c_3$ , including “*acute disease of eye*” and “*acute inflammatory disease of left eye*.” Note that the total number 35 can be obtained by multiplying 9 (8 new noun chunks and 1 original noun chunk for  $c_1$ ) by 4 (3 new noun chunks and 1 original noun chunk for  $c_3$ ), and subtracting 1 (the original concept name) from it.

#### 5.1.4 Identify missing hierarchical relations in source ontologies

In this step, we check if the newly generated concept names exist in the UMLS (i.e., exactly match the names of UMLS atoms) to identify potentially missing hierarchical relations between atoms in source ontologies. Given a concept name  $C$  (mapped to an atom  $AUI_C$ ) and a potential concept name  $S$  serving as its supertype, if the following conditions hold:

1.  $S$  can be mapped to a UMLS atom  $AUI_S$ ;
2.  $AUI_S$  comes from the same source ontology  $T$  as  $AUI_C$ ;
3. currently there is no hierarchical relation (either direct or indirect) between  $AUI_S$  and  $AUI_C$  claimed in  $T$ ; and
4.  $AUI_C$  has the same semantic type as  $AUI_S$ , or the set of semantic types of  $AUI_C$  contains that of  $AUI_S$  as a subset,

then we consider there is a potentially missing hierarchical relation between  $AUI_C$  and  $AUI_S$  (i.e.,  $AUI_C$  IS-A  $AUI_S$ ) in the ontology  $T$ . Note that missing hierarchical relations between atoms from different source ontologies are beyond the scope of this work. The semantic type requirement of  $C$  and  $S$  is to avoid ambiguities caused by concept names which may have multiple meanings. For example, the concept name “*cold*” could refer to lower temperature (with a semantic type “*Natural Phenomenon or Process*”) or a kind of disease (with a semantic type “*Disease or Syndrome*”).

For “*acute dacryoadenitis of left eye*” in Table 5.1, after the transformation, “*acute disease of eye*” is one of its potential new concept names which can be mapped to

atoms, while “*acute inflammatory disease of left eye*” cannot. By further mapping concept names to atoms, a potentially missing hierarchical relation between “*acute dacryoadenitis of left eye*” with AUI *A27761536* and “*acute disease of eye*” with AUI *A3463187* can be identified in the SNOMED CT.

It is worth noting that the potentially missing hierarchical relations identified by our method may contain redundancy. Here a missing hierarchical relation (say “ $x$  IS-A  $y$ ”) identified in an ontology  $T$  is considered redundant, if there exists another missing hierarchical relation “ $x$  IS-A  $z$ ” identified in  $T$  such that  $y$  is currently an ancestor of  $z$  in  $T$ . In this case, if “ $x$  IS-A  $z$ ” is a valid missing hierarchical relation, then “ $x$  IS-A  $y$ ” can be implied as valid by “ $x$  IS-A  $z$ ” and “ $z$  IS-A  $y$ .” Therefore, we further remove the potentially missing hierarchical relations that are redundant from the result.

## 5.2 Results

### 5.2.1 Identifying missing hierarchical relations

We applied our method to the English-language concept names in the UMLS (2019AB release). In total, our method identified 42,362 potentially missing hierarchical relations from 13 source ontologies in the UMLS. 39,359 out of 42,362 are non-redundant. Table 6.1 shows the number of potentially missing hierarchical relations (non-redundant) detected in each source ontology. Table 6.1 also presents each ontology’s size including the number of concepts and the number of direct hierarchical relations, as well as the number of existing hierarchical relations that can be identified by our transformation-based method. In total 149,568 existing hierarchical relations can be identified from 13 source ontologies, and 109,031 of them are direct hierarchical relations.

Among 39,359 potentially missing hierarchical relations identified, 36,997 were obtained from a single noun chunk replacement (1-replacement), 2,338 from two noun chunk replacements (2-replacement), and 24 from three noun chunk replacements (3-



replacement).

**Table 5.2:** The number of potentially missing hierarchical relations detected in the UMLS source ontologies in English, as well as the ontology size and the number of existing hierarchical relations that can be identified for each ontology.

| Source ontology | ontology size |                            | # of existing IS-A relations identified |        | # of potentially missing IS-A relations identified |
|-----------------|---------------|----------------------------|---|--------|--|
|                 | # of concepts | # of direct IS-A relations | direct + indirect                       | direct |  |
| MEDCIN          | 348,808       | 353,304                    | 30,001                                  | 23,692 | 16,779   |
| UWDA            | 61,127        | 62,285                     | 34,564                                  | 24,594 | 10,865   |
| FMA             | 102,595       | 104,341                    | 54,644                                  | 39,274 | 7,230  |
| SNOMEDCT_US     | 401,832       | 994,499                    | 19,859                                  | 14,529 | 3,833  |
| NCI             | 151,966       | 159,479                    | 688                                     | 539    | 334  |
| GO              | 49,907        | 77,067                     | 9,640                                   | 6,246  | 250  |
| SNOMEDCT_VET    | 36,527        | 40,689                     | 82                                      | 81     | 23   |
| HPO             | 16,222        | 18,313                     | 37                                      | 30     | 11   |
| CPM             | 3,079         | 3,853                      | 7                                       | 7      | 10   |
| UMD             | 27,309        | 12,889                     | 0                                       | 0      | 8  |
| PDQ             | 18,874        | 4,298                      | 43                                      | 36     | 8  |
| CPT             | 40,892        | 14,072                     | 1                                       | 1      | 7  |
| ATC             | 5,485         | 4,969                      | 2                                       | 2      | 1  |

## 5.2.2 Evaluation

To assess the effectiveness of our method for identifying missing hierarchical relations in the UMLS source ontologies, a sample of 200 hierarchical relations from SNOMED CT (the “*Clinical Finding*” subhierarchy) and a sample of 100 from the Gene Ontology were randomly selected and reviewed by domain experts. For each relation, we provided domain experts with the preferred names of the two concepts involved, as well as the links to the two concepts in their online browsers.

Domain experts verified that 173 out of 200 potentially missing hierarchical relations in the SNOMED CT (a precision of 86.5%) and 63 out of 100 in the Gene Ontology (a precision of 63%) are valid (i.e., true positives). Table 5.3 lists 15 valid examples.

**Table 5.3:** Examples of missing hierarchical relations confirmed by domain experts.

| Subtype concept  | Supertype concept  | Source ontology |
|--|--|-----------------|
| Abrasion and/or friction burn of buttock with infection (disorder)                                 | Superficial injury of buttock with infection (disorder)                              | SNOMEDCT_US     |
| Camptodactyly of right hand (disorder)   | Congenital deformity of right hand (disorder)  | SNOMEDCT_US     |
| Acute gastroduodenal ulcer with hemorrhage AND with perforation but without obstruction (disorder) | Peptic ulcer with hemorrhage AND with perforation but without obstruction (disorder) | SNOMEDCT_US     |
| Malignant melanoma of skin of forearm (disorder)   | Malignant neoplasm of skin of forearm (disorder)                                     | SNOMEDCT_US     |
| Deficiency of adenosylhomocysteinase (disorder)  | Deficiency of hydrolase (disorder)   | SNOMEDCT_US     |
| Infestation caused by Boophilus (disorder)   | Infestation caused by Ixodidae (disorder)  | SNOMEDCT_US     |
| Abscess of nasal septum (disorder)   | Inflammatory disorder of cartilage (disorder)  | SNOMEDCT_US     |
| Obsessive compulsive disorder caused by cocaine (disorder)   | Anxiety disorder caused by stimulant (disorder)                                      | SNOMEDCT_US     |
| Primary malignant neoplasm of frontal lobe (disorder)  | Malignant neoplasm of cerebral cortex (disorder)                                     | SNOMEDCT_US     |
| Rupture of anterior cruciate ligament of left knee (disorder)                                      | Injury of cruciate ligament of knee (disorder)                                       | SNOMEDCT_US     |
| negative regulation of testosterone biosynthetic process   | negative regulation of steroid hormone biosynthetic process                          | GO              |
| macrophage migration inhibitory factor binding   | enzyme binding   | GO              |
| response to camptothecin   | response to topoisomerase inhibitor  | GO              |
| formate dehydrogenase complex  | oxidoreductase complex   | GO              |
| negative regulation of transmembrane   | negative regulation of cellular process  | GO              |

Table 5.3 also contains four examples of missing hierarchical relations in SNOMED CT that were obtained by multiple noun chunk replacements. For instance, the missing hierarchical relation between “*Obsessive compulsive disorder caused by cocaine (disorder)*” and “*Anxiety disorder caused by stimulant (disorder)*” was obtained through the following two replacements: (1) “*Obsessive compulsive disorder*” IS-A “*Anxiety disorder*” in the NCI Thesaurus; and (2) “*Cocaine*” IS-A “*Psychostimulant*” and “*Psychostimulant*” IS-A “*Stimulant*” in the SNOMED CT.

### 5.2.3 Analyses of false positive cases

Based on the evaluation results from domain experts, we examined false positive cases (i.e., invalid missing hierarchical relations). More specifically, we looked into the noun chunks within the concept names and their replacement candidates (i.e., existing hierarchical relations) to find the potential causes.

Table 5.4 presents 7 invalid missing hierarchical relations as well as the existing hierarchical relations in the UMLS that were leveraged to obtain these invalid relations.

We noted that the main cause of false positives is that the biomedical meanings of replacement candidates are not considered to be more general than their corresponding noun chunks. This could relate to either incorrect existing hierarchical relations or different views of different ontologies. Take “*cellular response to beta-carotene*” IS-A “*cellular response to vitamin A*” detected in the Gene Ontology as an example. The domain experts believe that “beta-carotene” is an antioxidant that converts to vitamin A (which is not an hierarchical relation), while SNOMED CT has a hierarchical relation between “*Beta-carotene (substance)*” and “*Retinol (substance)*” (with a synonym “*Vitamin A*”), indicating that this is an incorrect hierarchical relation in the SNOMED CT. Consider “*Abscess of thumb of left hand (disorder)*” IS-A “*Abscess of finger of left hand (disorder)*” detected in the SNOMED CT. It was obtained by leveraging an existing relation “*Thumb*” IS-A “*Finger*” in both UWDA and FMA. However, the detected missing hierarchical relation is invalid, since in SNOMED CT “*Finger*” only includes the second to fifth digit of the hand (i.e., “*Thumb*” is not a “*Finger*”).

**Table 5.4:** Examples of false positives (or invalid missing hierarchical relations) and the existing hierarchical relations causing the false positives

| Subtype concept   | Supertype concept   | Source ontology | Existing IS-A relation(s) causing the false positive                                      |
|---|---|-----------------|---|
| Benign neoplasm of false vocal cord (disorder)                          | Benign neoplasm of vocal cord (disorder)                    | SNOMEDCT_US     | “false vocal cord” IS-A “vocal cord” in the NCI Thesaurus                                 |
| Deficiency of lysophospholipase (disorder)                              | Deficiency of triacylglycerol lipase (disorder)             | SNOMEDCT_US     | “lysophospholipase” IS-A “phospholipase” IS-A “triacylglycerol lipase” in the SNOMEDCT_US |
| Abscess of thumb of left hand (disorder)                                | Abscess of finger of left hand (disorder)                   | SNOMEDCT_US     | “thumb” IS-A “finger” in the UWDA and FMA   |
| Calculus of gallbladder with acute and chronic cholecystitis (disorder) | Calculus of gallbladder with acute cholecystitis (disorder) | SNOMEDCT_US     | “acute and chronic cholecystitis” IS-A “acute cholecystitis” in the MEDCIN                |
| cellular response to beta-carotene                                      | cellular response to vitamin A                              | GO              | “beta-carotene” IS-A “vitamin A” in the SNOMEDCT_US                                       |
| caprolactam metabolic process   | propylene metabolic process                                 | GO              | “caprolactam” IS-A “propylene” in the SNOMEDCT_US   |
| cellular response to ammonium ion                                       | cellular response to ammonia                                | GO              | “ammonium ion” IS-A “ammonia” in the SNOMEDCT_US  |

#### 5.2.4 Effect of restricting the hierarchical source for noun chunk replacement

Relating to the subtle ontology difference, a natural question is whether restricting the hierarchical relations leveraged for noun chunk replacement to be in the same ontology will have an effect on the performance of our method. To study this, we performed an experiment by restricting replacement candidates in the same ontology, which resulted in a total of 20,754 potentially missing hierarchical relations, compared to 39,359 without applying the restriction.

We further looked into the evaluated samples regarding the performance comparison. For SNOMED CT, 173 out of 200 evaluated relations (without applying the restriction) are valid, achieving a precision of 86.5%. Among 200 evaluated ones, 107 of them can be obtained by applying the restriction, and 103 out of 107 are valid, achieving a precision of 96.26%. Therefore, the precision is increased by 9.76% with the restriction applied. However, the number of valid missing hierarchical relations is decreased from 173 to 103, a 40.46% reduction. For Gene Ontology, 63 out of 100 evaluated relations (without applying the restriction) are valid, achieving a precision of 63%. Among 100 evaluated ones, 21 of them can be obtained by applying the restriction, and 18 out of 21 are valid, achieving a precision of 85.71%. Therefore, the precision is increased by 22.71%. However, the number of valid missing hierarchical relations is decreased from 63 to 18, a 71.43% reduction. It can be seen that although restricting to the same source ontology for noun chunk replacement can improve the precision to some extent, leveraging multiple sources can identify more missing hierarchical relations to a greater extent while still achieving acceptable precision.

### 5.3 Discussion

In this work, we introduced a transformation-based method to replace noun chunks in a concept name with more general concept names in order to detect potentially missing hierarchical relations in the UMLS source ontologies. To find noun chunk replacement, we leverage abundant knowledge of hierarchical relations between concept names provided by the UMLS.

#### 5.3.1 Distinction with related work

Other auditing methods designed for a specific ontology including pattern-based, lexical-based and deep learning-based methods usually rely on the knowledge in the ontology itself and require transferring knowledge to features for representing concepts in order to identify missing hierarchical relations between concepts [22, 23, 25–27, 31, 32, 43]. Therefore, the effectiveness of such methods to some extent relies on the ontology itself (i.e., internal knowledge), while our method leverages both internal and external knowledge through the UMLS to perform the auditing. More importantly, our method enables the auditing of multiple source ontologies at the same time.

In addition, unlike previous related work on auditing the UMLS that mainly focused on auditing high level views (e.g., semantic types, concepts/CUIs, relations between concepts), this work intends to audit the UMLS source ontologies at the atom level.

#### 5.3.2 Exact versus normalized matching

For parsing and mapping concept names, we directly used the exact names without performing any normalization. We further tried normalizing concept names (after noun chunks were identified) using the UMLS lexical tool LuiNorm [105]. We also utilized the normalized format for generating replacement candidates for noun chunks

and mapping newly constructed concept names to atoms. As a result, the potentially missing hierarchical relations identified using normalized matching contain all the 39,359 ones identified by exact matching as a subset. In addition, the normalized matching identified 10,627 extra potentially missing hierarchical relations.

Indeed, normalized matching helped identify extra valid missing hierarchical relations. For example, a missing hierarchical relation between “*Malignant neoplasm of connective tissue*” and “*Neoplasm of connective tissues*” in the SNOMED CT was identified by normalized matching, since “*tissues*” was normalized to “*tissue*.” However, there were also invalid cases identified. For instance, “*Asymmetry*” is a child of “*Symmetries*” in the SNOMED CT. Performing normalization resulted in “*Asymmetry*” IS-A “*Symmetry*” and thus an invalid missing hierarchical relation: “*Asymmetry of mandible*” IS-A “*Symmetry of mandible*.” Since the main focus of this work is the transformation-based method, it is beyond the scope of this work to thoroughly compare the actual performances of the exact matching and normalized matching, as it requires additional manual evaluation by domain experts.

### 5.3.3 Potential for concept enrichment

Since our focus in this work is to identify missing hierarchical relations in the UMLS source ontologies, we require that the two atoms involved in a potentially missing hierarchical relation be from the same ontology. For those ones with the two atoms coming from different source ontologies, missing concepts may be identified for concept enrichment in source ontologies. That is, if the supertype atom does not appear in the same source ontology as the subtype atom, then the supertype atom may be a potentially missing concept (i.e., new concept) for the ontology or a missing synonym for an existing concept in the source ontology of the subtype atom.

### 5.3.4 Applicability to a specific ontology

Although our method was designed for auditing multiple source ontologies in the UMLS, it can be applied within a specific ontology such as the SNOMED CT itself without using the UMLS. A question that may arise is: Will this give the same results obtained for restricting the hierarchical relations leveraged for noun chunk replacement to be in the UMLS SNOMED CT? The answer to this question depends on whether the hierarchical relations in the UMLS SNOMED CT are identical to that in the original SNOMED CT. It is worth noting that relations in the UMLS are expressed in terms of CUIs (concepts) and AUIs (atoms or concept names). For the September 2019 release of SNOMED CT (US edition) integrated in the UMLS (2019AB release), only hierarchical relations between designated preferred names of SNOMED CT concepts are maintained. Therefore, if we only leverage such hierarchical relations between preferred names of concepts when applying our method within the SNOMED CT, then the same results will be obtained; however, if we leverage additional hierarchical relations such as those between synonyms of concepts, then more results will be obtained and need further domain expert evaluation.

## 5.4 Conclusion

In this Chapter, a concept name transformation-based auditing method is introduced to detect potentially missing hierarchical relations in the UMLS source ontologies. Leveraging rich knowledge in the UMLS (2019AB release), our method is able to audit multiple ontologies at the same time. Experts' evaluation showed the effectiveness of our method (a precision of 86.5% for SNOMED CT and 63% for the Gene Ontology). Further analyses of invalid missing hierarchical relations derived by our method revealed additional quality issues in the source ontologies.

## CHAPTER 6. A Lexical-based Formal Concept Analysis Method to Identify Missing Concepts in the NCI Thesaurus

As part of the ontology evolution process, new concepts are regularly added in response to the evolving domain knowledge and emerging applications. Most existing concept enrichment methods suggest new concepts via directly importing knowledge from external sources. In this chapter, we introduced an FCA-based method that utilizes the intrinsic knowledge within the ontology itself. Compared with the traditional FCA-based methods which take logical definitions as attributes (i.e., difficult to validate new concepts), our approach considers lexical features (i.e., words appearing in the concept names) as FCA attributes while generating formal context. As a result, formalizing new concepts also brings bags of words that could be used to name the concepts which are more convenient to validate compared with sets of logical definitions.

### 6.1 Method

This method mainly consists of two steps: (1) pre-processing concept names and constructing FCA formal context; and (2) performing FCA via a multistage intersection algorithm to identify potentially missing (or new) concepts in the NCI Thesaurus.

#### 6.1.1 Constructing formal context

Given a collection of concepts in the ontology, we consider all the concepts as FCA objects  $O$  and words appearing in the concept names (i.e., lexical features) as FCA attributes  $A$ , respectively. With the binary relation  $R \subseteq O \times A$  specifying whether concept  $o \in O$  contains word  $a \in A$ , we can construct the FCA formal context  $K = (O, A, R)$ .

Since words appearing in concept names may have variations (e.g., plural vs. sin-



gular forms) or synonyms, we perform attribute/word normalization to create a more robust FCA formal context. For word variations, we normalize words appearing in concept names using LuiNorm [105], a lexical tool provided by the UMLS. For example, “bones” can be normalized to “bone”. Regarding word synonyms, we leverage concepts in the NCI Thesaurus with single-word preferred names and single-word synonyms. More specifically, if a word  $w$  itself is the preferred name of an NCI Thesaurus concept and has a synonym  $s$  that is also a single word, then we maintain a mapping between the synonym  $s$  and the preferred name  $w$ . This way words with the same meanings can be normalized to their preferred names thus the same attribute.

### 6.1.2 Identifying potentially missing concepts

To derive FCA formal concepts, we leverage the idea of a faster concept analysis algorithm introduced in [106], which is to perform multistage intersection on each pair of formal concepts from the initial formal concept set consisting of all objects, until no more new formal concept is generated. The pseudocode of the algorithm is shown in Fig. 6.1.

---

#### Algorithm 1 Identifying Missing Concepts

---

```

1: Input: Formal context  $(O, A, R)$ 
2: Output: Missing concept set  $M$ 
3: Initialization:
4: Original set  $S_0 \leftarrow \{o^\uparrow | o \in O\}$ 
5: Initial set  $I \leftarrow S_0$ 
6: Newly derived formal concept set  $N \leftarrow S_0$ 
7: while  $N \neq \emptyset$ 
8:   Last iteration formal concept set  $L \leftarrow I$ 
9:   for each pair  $(C_x, C_y)$  in  $L$ 
10:     $I.add(\text{Intersection}(C_x, C_y))$ 
11:    $N \leftarrow (I - L)$ 
12:  $M \leftarrow I - S_0$ 

```

---

**Figure 6.1:** Pseudocode of identifying potentially missing concepts by multistage intersection.

In practice, for computation convenience, we perform operations on the lexical

feature sets (i.e., using FCA attribute sets to represent FCA formal concepts). The initial set of FCA formal concepts is a set of FCA attribute sets, that is, the lexical feature sets of all the original concepts (i.e.,  $\{o^\uparrow \mid o \in O\}$ ). In the first iteration, we compute the intersection of each pair of FCA attribute sets in the initial set; and if the result is not included in the initial set, we add it into the initial set. We repeat this process until no new FCA attribute set can be derived. Each newly generated FCA attribute set is taken as the lexical feature set of a potentially missing concept among the given concepts. An advantage of using lexical features (or words) as FCA attribute sets is that these words can be further leverage to name the newly discovered concepts.

### 6.1.3 Illustrative example

Fig. 6.2 shows a simple example of FCA formal context in a tabular format generated from the concept *Breast Fibroepithelial Neoplasm* (*C40405*) and its descendants in the NCI Thesaurus. The cells with check marks represent the binary relation between the concepts and their lexical features. Note that word “Tumor” is normalized to “neoplasm,” since it is a synonym of *Neoplasm* (*C3262*) in the NCI Thesaurus.

|   | juvenile | fibroepithelial | malignant | breast | fibroadenoma | neoplasm | complex | borderline | pericanalicular | intracanalicular | benign | phyllode | giant |
|---|----------|-----------------|-----------|--------|--------------|----------|---------|------------|-----------------|------------------|--------|----------|-------|
| C3744: Breast Fibroadenoma                  |          |                 |           | ✓      | ✓            |          |         |            |                 |                  |        |          |       |
| C7575: Breast Phyllodes Tumor               |          |                 |           | ✓      |              | ✓        |         |            |                 |                  |        | ✓        |       |
| C4271: Breast Intracanalicular Fibroadenoma |          |                 |           | ✓      | ✓            |          |         |            |                 | ✓                |        |          |       |
| C4272: Breast Pericanalicular Fibroadenoma  |          |                 |           | ✓      | ✓            |          |         |            | ✓               |                  |        |          |       |
| C4273: Breast Giant Fibroadenoma            |          |                 |           | ✓      | ✓            |          |         |            |                 |                  |        |          | ✓     |
| C4276: Breast Juvenile Fibroadenoma         | ✓        |                 |           | ✓      | ✓            |          |         |            |                 |                  |        |          |       |
| C5194: Breast Complex Fibroadenoma          |          |                 |           | ✓      | ✓            |          | ✓       |            |                 |                  |        |          |       |
| C4504: Malignant Breast Phyllodes Tumor     |          |                 | ✓         | ✓      |              | ✓        |         |            |                 |                  |        | ✓        |       |
| C5196: Benign Breast Phyllodes Tumor        |          |                 |           | ✓      |              | ✓        |         |            |                 |                  | ✓      | ✓        |       |
| C5316: Borderline Breast Phyllodes Tumor    |          |                 |           | ✓      |              | ✓        |         | ✓          |                 |                  |        | ✓        |       |
| C40405: Breast Fibroepithelial Neoplasm     |          | ✓               |           | ✓      |              | ✓        |         |            |                 |                  |        |          |       |

**Figure 6.2:** An example of FCA formal context generated by the concept *Breast Fibroepithelial Neoplasm* (*C40405*) in the NCI Thesaurus and its descendants in company with their lexical features. Word “Tumor” is normalized to “neoplasm” and word “Phyllodes” is normalized to “phyllode.” An FCA formal concept (marked by blue cells) with FCA attribute set {breast, neoplasm} is considered as a potentially missing concept among the given concepts.

Given the FCA formal context, the FCA formal concept with attribute set {breast, neoplasm} (see blue cells in Fig. 6.2) can be derived by intersecting the attribute sets

of *Borderline Breast Phyllodes Tumor* (*C5316*) and *Breast Fibroepithelial Neoplasm* (*C40405*). Therefore, a concept with lexical feature set {breast, neoplasm} is considered as a potentially missing concept for the given FCA formal context. This example only intends to illustrate how our method works, and one may have noticed that *Breast Neoplasm* (*C2910*) is an existing concept in the NCI Thesaurus although it is not among the given concepts. For the actual implementation of our method, we further check if the newly generated concepts are existing in the NCI Thesaurus and ensure the removal of such cases from the list of potentially missing concepts.

## 6.2 Results

### 6.2.1 Summary result

We applied our method to the sub-hierarchies under *Disease or Disorder* (*C2991*) in the 19.08d version of the NCI Thesaurus. Table 6.1 shows the numbers of existing concepts, newly generated concepts, and potentially missing concepts respectively for each sub-hierarchy. For example, there are 10,996 existing concepts in the *Neoplasm* (*C3262*) sub-hierarchy; and FCA generated a total of 8,511 new concepts, among which 7,737 were potentially missing concepts in the NCI Thesaurus.

Note that potentially missing concepts are detected in terms of the given FCA formal context (or the given collection of the input concepts). Therefore, the missing concepts detected in a sub-hierarchy may overlap with those detected in another sub-hierarchy. In total, 8,983 unique potentially missing concepts were identified among these sub-hierarchies.

### 6.2.2 Preliminary evaluation

We performed a preliminary evaluation to validate the potentially missing concepts identified using the external knowledge in the UMLS which integrates millions of biomedical concepts from more than 200 source ontologies [87].

**Table 6.1:** The numbers of existing concepts, newly generated concepts, potentially missing concepts, and missing concepts validated via UMLS for each sub-hierarchy under *Disease or Disorder (C2991)*.

| Sub-hierarchy                      | # of Concepts | # of Newly Generated Concepts |                          |                         |
|------------------------------------|---------------|-------------------------------|--------------------------|-------------------------|
|                                    |               | Total                         | # of Potentially Missing | # of Validated via UMLS |
| C27551: Disorder by Site           | 13,595        | 9,114                         | 7,864                    | 451                     |
| C3262: Neoplasm                    | 10,996        | 8,511                         | 7,737                    | 289                     |
| C53529: Non-Neoplastic Disorder    | 4,198         | 1,279                         | 813                      | 227                     |
| C8278: Cancer-Related Condition    | 578           | 491                           | 374                      | 28                      |
| C4873: Rare Disorder               | 915           | 283                           | 196                      | 44                      |
| C89328: Pediatric Disorder         | 528           | 280                           | 218                      | 20                      |
| C28193: Syndrome                   | 907           | 266                           | 204                      | 31                      |
| C3101: Genetic Disorder            | 159           | 52                            | 30                       | 6                       |
| C2893: Psychiatric Disorder        | 231           | 45                            | 29                       | 11                      |
| C3113: Hyperplasia                 | 81            | 24                            | 17                       | 8                       |
| C3340: Polyp                       | 110           | 24                            | 7                        | 2                       |
| C35470: Behavioral Disorder        | 49            | 19                            | 9                        | 0                       |
| C3075: Hamartoma                   | 63            | 15                            | 6                        | 0                       |
| C26684: Radiation-Induced Disorder | 25            | 5                             | 3                        | 0                       |

For each potentially missing concept identified, we checked whether its lexical feature set can be matched to any concept name from the external ontologies in the UMLS. We found 592 out of 8,983 potentially missing concepts are included in the external ontologies in UMLS (see Table 6.1 for the number of missing concepts validated via UMLS for each sub-hierarchy). Table 6.2 lists 10 examples of validated missing concepts (in the form of lexical feature sets) and matched concept names in the UMLS ontologies.

Since a matching concept may be from multiple UMLS ontologies, we further looked into the ontologies that contributed most to the validation of the 592 identified potentially missing concepts. The top 10 in terms of the number of matched concepts (in parentheses) are listed as follows: Consumer Health Vocabulary - CHV (328), SNOMED CT US Edition - SNOMEDCT\_US (245), Read Codes - RCD (135), MedDRA - MDR (124), ICPC2-ICD10 Thesaurus - ICPC2ICD10ENG (101), MSH (97), Metathesaurus Names - MTH (79), MEDCIN (78), Online Mendelian Inheri-

tance in Man - OMIM (55), and Logical Observation Identifiers Names and Codes - LNC (52).

**Table 6.2:** Ten examples of validated missing concepts and their matched concepts in the UMLS ontologies.

| Lexical Feature Set of Missing Concept     | Matched Concept (External ontology)            |
|--|--|
| {carcinoma, papillary, urothelial}         | Papillary urothelial carcinoma (SNOMEDCT_US)   |
| {borderline, serous, tumor}                | Serous borderline tumor (SNOMEDCT_US)          |
| {intestinal, lymphoma}                     | Intestinal lymphoma (SNOMEDCT_US)              |
| {adrenal, carcinoma}                       | Adrenal carcinoma (OMIM)                       |
| {in, breast, carcinoma, situ}              | breast carcinoma in situ (CHV)                 |
| {fossa, piriform}                          | Piriform Fossa (MSH)                           |
| {cellular, pigmentation}                   | cellular pigmentation (GO)                     |
| {b-cell, cutaneous, lymphoma, primary}     | Primary cutaneous B-cell lymphoma (MEDCIN)     |
| {gastric, sarcoma}                         | gastric sarcoma (MEDCIN)                       |
| {adenocarcinoma, breast, metaplasia, with} | breast adenocarcinoma with metaplasia (MEDCIN) |

### 6.3 Discussion

In this work, we leveraged words in concept names and FCA to detect potentially missing concepts in the NCI Thesaurus. The preliminary evaluation via UMLS-based validation indicates that our method has the potential to identify missing concepts for concept enrichment of the NCI Thesaurus. However, there are still several things that need our attention.

First, the potentially missing concepts detected by our method may not be directly imported into an ontology. This is because different ontologies are developed for disparate purposes and have varying target applications, and a concept that is essential for an ontology may not be necessary for another. Further reviews and evaluations by the ontology curators are still required to decide whether a concept is meaningful and should be added according to the scope of the ontology and its potential applications.

Another thing is that the “subconcept-superconcept” relations between formal

concepts derived from lexical features could be different from the hierarchical IS-A relations in the original ontology. For instance, *Breast Neoplasm* and *Breast* are two new concepts generated based on the FCA formal context in Fig. 6.2. Although the two concepts have a “subconcept-superconcept” relation in terms of the FCA word attributes, they do not form a valid IS-A relation. In fact, *Breast* locates in a different sub-hierarchy *Organ*. A potential solution to avoid such cases is to use enriched lexical features for a concept, which includes its ancestor’s lexical features. This way, the original hierarchical relation will be captured in the initial FCA formal context, and thus the new concepts generated by attribute set intersection will locate within the same sub-hierarchy with the root concept. However, the enriched lexical features may make it more difficult to decide which words to use for naming a concept. To deal with this, we could further leverage both logical definitions and lexical features to identify and name missing concepts.

## 6.4 Conclusion

In this chapter, we present a lexical- and FCA-based method that utilizes intrinsic knowledge of an ontology to detect potentially missing concepts. We applied our method to the NCI Thesaurus *Disease or Disorder* sub-hierarchy and identified 8,983 potentially missing concepts. The preliminary evaluation via external validation using UMLS showed encouraging evidence for the effectiveness of our method.

## CHAPTER 7. Exploring Deep Learning-based Approaches for Predicting Concept Names in SNOMED CT

Although many automatic methods have been proposed to identify missing or new concepts in biomedical ontologies, proper naming of those new concepts remains challenging and relies on the curators of biomedical ontologies. However, it is hard for different curators to maintain the same standard and keep consistent while naming thousands of concepts. Also, according to the experiment results from Zhu et al’s work [19], even in well-constructed and mature ontologies such as SNOMED CT, there still exists a large number of missing concepts. It is labor-intensive and time-consuming for curators to manually find appropriate and unambiguous names for a large number of concepts. Therefore, automated methods are highly desirable to provide suggestions on concept names to reduce curators’ manual burden and accelerate the ontology maintenance process.

By using the method introduced in Chapter 6 and the method in our previous work [22], we are able to generate a bag of words that are necessary to construct the name for a potentially missing concept. However, the words may be unordered, or the order of the words may not be consistent with the naming convention of the given ontology. For example, we could get bag of words  $\{of, neoplasm, malignant, upper, lobe, right, of, lung\}$  whose proper name should be “*malignant neoplasm of right upper lobe of lung.*”

In this Chapter, based on our previous work, we further try to generate proper sequence order (i.e., concept names) for a given bag of words. We explore three deep learning-based approaches to automatically predict concept names that comply with the naming convention of SNOMED CT. These deep learning models are simple neural network, Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN) combined with LSTM.

## 7.1 Method

In this work, we focus on addressing the problem of predicting the word sequence given a bag of words for naming a concept. To achieve this, we divide this task into the following two subtasks. Firstly, given a word sequence, we train the models to determine whether the given sequence or order is correct (meaningful and satisfying the naming convention of SNOMED CT) or not (binary classification). Secondly, given a bag of words, we utilize the trained models to predict its correct word sequence. Since the trained models return different confidence levels of the correctness judgment for different sequences, we test all the possible sequences of the given words and choose the one(s) with the highest confidence as the predicted concept name(s). Further, we implement a two-step filter to eliminate those potential incorrect candidate concept name(s).

### 7.1.1 Word embedding & data preprocessing

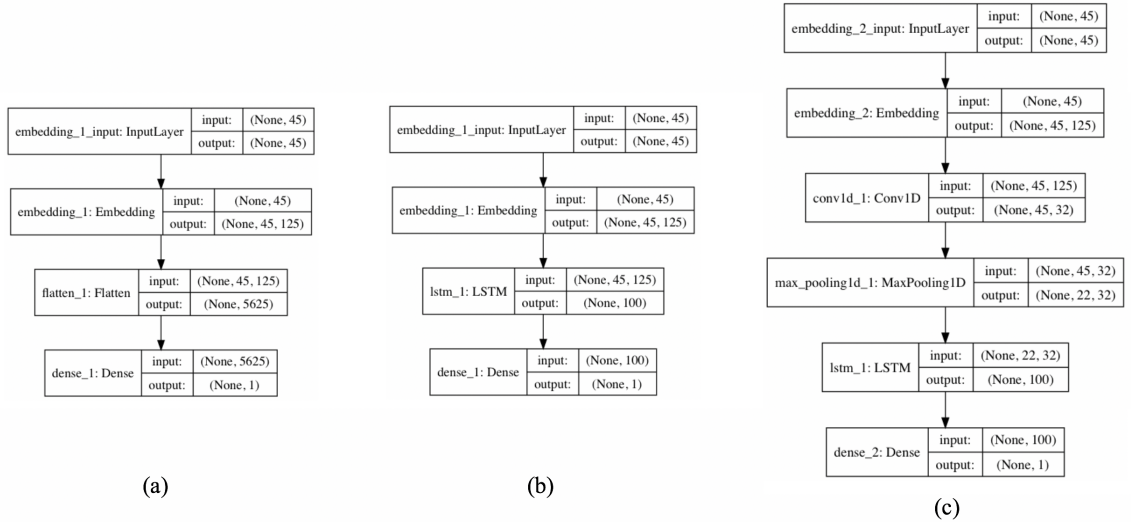
We first use Word2Vec [107] provided by Gensim [108] to learn vector representation of words in SNOMED CT concept names. Each word is mapped to a 125-length vector (word embedding) based on the word(s) surrounding it, and thus mapped to a higher dimensional space. The mappings between words and their word embeddings are stored in a matrix which later will be reused in training the neural network models.

To train the models, we need both positive and negative training data. The original sequences of words (concept names) are labeled as “1” or correct. Then, for each correct sequence, we generate  $n$  incorrect sequences ( $n = 5$  in this work) by randomly disordering the sequence and the generated sequences are labeled as “0” or wrong. Since concept names in SNOMED CT are in different lengths (ranging from 1 to 39), all sequences are padded to the same length of 45.



### 7.1.2 Neural networks for classifying word sequences

We use Keras [109], a high-level neural networks API, to implement three types of deep learning models as shown in Fig. 7.1. For all three models, there is an embedding layer right after the input layer. It is responsible for mapping each word in the input sequence to its word embedding. The pre-computed word embedding matrix will serve as the trained weight for this layer.



**Figure 7.1:** Three neural network models used for the classification task. (a) is a simple neural network for binary classification; (b) is an LSTM neural network; (c) is the combination of convolutional neural network and LSTM. “None” means that the dimension is variable.

The first model (Fig. 7.1(a)) is a simple neural network for binary classification. Since word embedding increases the input dimension, we need to flatten the input. Then we use a dense layer that has a *sigmoid* activation function to generate the prediction result. For the classification task, we use *binary\_crossentropy* in Keras, which is often used for binary classification problems, in all three models while training.

The concept names in SNOMED CT are in various lengths, and the position of a word may not only depend on the words next to it. This requires the model to be able to learn long-term context and dependencies between words in the input sequence. LSTM is proven to be good at learning such dependencies [67]. Thus we also adopt

LSTM (Fig. 7.1(b)) for binary classification that contains a single LSTM layer with 100 LSTM memory unit in the middle.

When it comes to determining the correctness of a sequence, there may exist featured patterns which can indicate whether a sequence order is correct or not. Thus, to capture those potential featured patterns, we employed a model in which convolutional neural network (CNN) is adopted and combined with LSTM (Fig. 7.1(c)). CNN is commonly applied to analyzing visual imagery and it can identify lower level features from the minimum unit of the input which eventually may improve the classification process [110]. CNN also benefits sentence classification, such as sentiment analysis and question classification [111]. Regarding our work, the input is sequence of words (concept name) which can be considered as one dimension spatially. By using one-dimension CNN (Conv1D), certain word combinations or patterns will be selected as lower level feature and these learned spatial features will then be learned by an LSTM layer. In this model, the number of output filters in the Conv1D layer is 32 and the window size is set to three. The *pool\_size* [112] for the *MaxPooling1D* layer is two.

As mentioned in the data preprocessing step, the number of input data labeled as “1” is less than those labeled as “0,” with a proportion of 1 : 5. To lessen the impact of unbalanced data, we assign different weights to different classes so that during training, the model will weight class “1” more when adjusting the weight.

### 7.1.3 Predicting concept names given bags of words

To suggest candidate concept name for a given bag of words, we first generate all its permutations (i.e., all possible sequences). Then we use the trained models to classify those generated sequences to check which one is valid. While performing classification, the neural network models could return a confidence score (probability) for a sequence to be valid. Thus, we select the sequence(s) with the highest confidence score to be

valid as the potential concept name(s) for a given bag of words. Because the number of permutations of  $n$  distinct objects is  $n$  factorial. When  $n$  becomes relatively large, the computation time increases dramatically. Thus, in this work, we only provide prediction for concept names whose length is less than or equal to 9.

Since there may exist multiple sequences with the highest confidence score for a given bag of words, we implement a two-step filtering process to further select the “best” candidate(s). In the first step, we leverage the idea of Viterbi algorithm [113] which returns the most likely sequence of hidden states, to remove those “invalid” ones. It is based on the assumption that if a word  $A$  has never been placed before (or after) another word  $B$  in the training data set, then  $A$  is not likely to appear before (or after)  $B$  in the candidate concept names. In the second step, we leverage similar concept names to further reduce the candidate list. For a bag of words, we first find the most similar concept name (in terms of the bag of words) in the training data, where similarity is calculated by dividing the number of words that appear in both bags of words by the total number of distinct words in two bags. Then we utilize “Levenshtein distance” [114] to compute the editing distances between each candidate name and the most similar concept name, and the one with the least distance will be selected as the “best” candidate concept names.

## 7.2 Experiment & result

To validate the effectiveness of our method, we focused on two research questions:

1. Are the deep learning models able to determine if a sequence of words is a valid concept name in SNOMED CT? (binary classification)
2. Given a bag of words, can our method generate the correct sequence using those words and thus provide suggestions on how to name a concept in SNOMED CT? (sequence prediction)

### 7.2.1 Experiment setup

To explore the first question regarding binary classification, we performed two experiments. In the first experiment, we randomly separated the concept names in the March 2018 US Edition of SNOMED CT into two groups: training and test datasets. There are 1,753,513 labeled sequences in the training data set and 1,784,744 labeled sequences in the test data set. In the second experiment, we trained our models by all the concept names in the September 2017 US Edition of SNOMED CT (along with randomly disordered ones). For testing, we extracted all the new concept names that were added into the March 2018 US Edition of SNOMED CT and generated disordered ones. For each of these two experiments, after training our models with the training dataset, we evaluated our models using the test dataset.

For the second question regarding sequence prediction, we evaluated our method in two ways. In the first way, we considered the sequence orders of the newly added concept names in the March 2018 US Edition of SNOMED CT as the ground truth for sequence prediction. For each concept name, we regarded it as a bag of words and used our method to generate candidate concept names. Then we compared these suggestions with the ground truth. In the second way, we identified a collection of missing concepts (in the September 2017 US Edition of SNOMED CT) and the corresponding bags of words that are necessary to construct their names using Cui et al.’s method [22]. A total of 60 concepts in the form of bags of words were obtained for testing the performance of word sequence prediction with the help of a human annotator for validation.

### 7.2.2 Result for binary classification

In the first experiment for binary classification, we first tested different thresholds of confidence (for a sequence to be labeled as positive) for three models to achieve the best F1 score. The results are 0.5 for simple neural network, 0.8 for LSTM,

and 0.7 for CNN and LSTM, respectively. The results of binary classification for three models using these thresholds are shown in Table 7.1, where it can be seen that the LSTM model outperformed the other two models, and achieved the best performance: an accuracy of 94.72%, a precision of 84.59%, a recall of 83.51%, an F1 score of 84.05%, and an FP-rate of 3.04%. The simple neural network performed the worst, and generated more false positives than the other two models. This is not surprising because while determining if a sequence is correct or not, the order of words matters and the simple neural network may not be able to learn long-term dependencies. In addition, it is shown that the combination of CNN and LSTM did not improve the performance of binary classification.

**Table 7.1:** Result of binary classification for Experiment I.

|           | Simple Neural Network | CNN and LSTM | LSTM   |
|-----------|-----------------------|--------------|--------|
| Accuracy  | 73.58%                | 92.15%       | 94.72% |
| Precision | 35.42%                | 74.23%       | 84.59% |
| Recall    | 71.08%                | 81.03%       | 83.51% |
| F1 Score  | 47.28%                | 77.48%       | 84.05% |
| FP-rate   | 25.91%                | 5.63%        | 3.04%  |

The result of the second experiment for binary classification is shown in Table 7.2. It can be seen that the LSTM model still achieved the best performance. The main difference between the two experiments is that the first one’s evaluation is within the same version of SNOMED CT, however, the second one is using concept names from a newer version of SNOMED CT to test the model which was trained by the older version. Although the training data in the second experiment is much more than the testing data, it exhibited a similar performance as the first experiment.

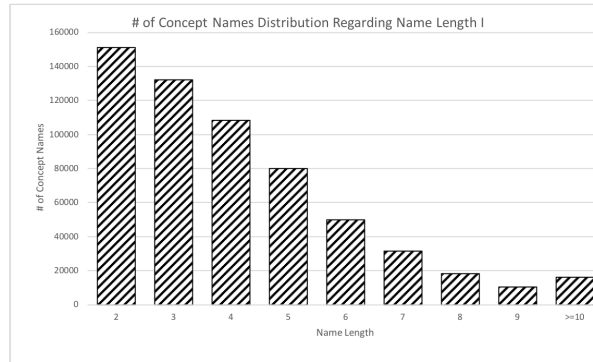
Since the LSTM model achieved the best performance in the binary classification subtask, we utilized it to perform sequence prediction for given bags of words in the next step.

**Table 7.2:** Result of binary classification for Experiment II.

|           | Simple Neural Network | CNN and LSTM | LSTM   |
|-----------|-----------------------|--------------|--------|
| Accuracy  | 71.75%                | 93.35%       | 95.05% |
| Precision | 31.67%                | 81.63%       | 87.56% |
| Recall    | 60.05%                | 77.56%       | 81.97% |
| F1 Score  | 41.47%                | 79.54%       | 84.67% |
| FP-rate   | 25.90%                | 3.49%        | 2.33%  |

### 7.2.3 Result for sequence prediction

For the sequence prediction subtask, we first used the trained LSTM model to predict correct sequence orders for the newly added concept names in the March 2018 US Edition of SNOMED CT. Given the computational challenge for testing all possible sequences and the fact that less than 5% of concept names in SNOMED CT are in the length of more than or equal to ten (see Fig. 7.2 for the distribution of concept names in terms of their lengths), we performed the sequence prediction for concepts whose lengths are less than ten.



**Figure 7.2:** Number of concept names in terms of the name length for all concepts in the March 2018 US Edition of SNOMED CT.

For a bag of words, our method will provide a set of candidate concept names. If a candidate name is in correct order (i.e., a valid name), then it is considered as a true positive case; otherwise, it is considered as a false positive case. If the correct sequence is not included in the set of candidate concept names, we have one false negative case.

Table 7.3 shows the performance of our LSTM-based sequence prediction approach for predicting concept names with an overall F1 score of 63.41%. It can be seen that concepts whose names are in the length of two and three received an F1 score of above 80%, concepts in length of four and five received an F1 score of above 60%, concepts in length of six and seven received an F1 score of above 50%, concepts in length of eight received an F1 score of 49.1%, and concepts in length of nine received an F1 score of 39.91%. This indicates that as the length of concept names grows, more sequences (false positive cases) might be included in the candidate set which leads to decreasing performance. However, even for concept names with length of six, seven or eight, the performance is still acceptable. Overall, the F1 score of our model is 63.41%.

**Table 7.3:** Result of LSTM-based sequence prediction in terms of the length of concept names. Training data is from September 2017 US Edition of SNOMED CT and test data is the newly added concepts in the March 2018 Edition.

| Length of Concept Name    | 2      | 3      | 4      | 5      | 6      | 7      | 8      | 9      | All    |
|---------------------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| Number of True Positives  | 735    | 1274   | 909    | 918    | 565    | 446    | 316    | 184    | 5347   |
| Number of False Positives | 154    | 310    | 556    | 465    | 375    | 278    | 287    | 332    | 2757   |
| Number of False Negatives | 155    | 321    | 590    | 665    | 627    | 465    | 368    | 222    | 3413   |
| Precision                 | 82.68% | 80.43% | 62.05% | 66.38% | 60.11% | 61.60% | 52.40% | 35.66% | 65.98% |
| Recall                    | 82.58% | 79.87% | 60.64% | 57.99% | 47.40% | 48.96% | 46.20% | 45.32% | 61.04% |
| F1 Score                  | 82.63% | 80.15% | 61.34% | 61.90% | 53.00% | 54.56% | 49.10% | 39.91% | 63.41% |

We also performed another way to evaluate the performance of our LSTM-based sequence prediction approach, by utilizing Cui et al.’s method in [22] to generate bags of words that are necessary to construct the names of new concepts. We obtained 60 bags of words and performed name prediction for them. For each predicted name, a human annotator manually examined whether the generated sequence order is correct and conforming to the name convention of SNOMED CT. Among 60 cases, 44 out of them were considered as correct by the human annotator. Table 7.4 shows our LSTM-based approach achieved an F1 score of 73.95%. The positive examples include “*malignant neoplasm of blood vessel of thorax*,” “*structure of layer of tunica vaginalis*” and “*open wound of limb without complication*.” This indicates that the LSTM-based

method can be applied for naming a new concept based on the bags of words featuring a concept.

**Table 7.4:** Result of LSTM-based sequence prediction for names of missing concepts identified by Cui et al.’s method in [22].

|                           |        |
|---------------------------|--------|
| Number of Concepts Names  | 60     |
| Number of True Positives  | 44     |
| Number of False Positives | 15     |
| Number of False Negatives | 16     |
| Precision                 | 74.58% |
| Recall                    | 73.33% |
| F1 Score                  | 73.95% |

### 7.3 Discussion

#### 7.3.1 Potential Factors Affecting the Prediction Performance

There are mainly two factors that potentially affect the performance of our prediction model – the size of bags of words and the words in the bag.

For the first factor, as the length of concept names increases, it becomes more difficult for the model to predict the correct sequence. This is because when the size of a bag of words increases, the number of sequences that need to be classified by the model increases dramatically. Since we assume that there is only one correct concept name for a bag of words, the false positive cases may greatly lower the model’s precision. For instance, for a bag of words of size five without duplicate words, it has 120 permutations and only one of them is correct. If we only have two false positive cases, the accuracy is about 98%, however, the actual precision is only 33% and F1 measure is 49%. Thus, the length of the concept name is an important factor affecting the performance of model. In our experiments, 50% of the testing data are of a length that is larger than or equal to five.

The second factor that may affect the model’s performance involves the words in



the bag. Our model may generate multiple candidate concept names for a given bag of words. This may be because that some words in the bag have the same role in other concept names. For example, for the concept name “*ultrasound guided biopsy of left and right breast*,” our model may be confused about the order of words “*left*” and “*right*,” because they may appear separately in other concepts but with the same roles or patterns. Another typical case is related to duplicate words in a bag. If a bag contains two (or more) identical words (e.g. multiple “*of*” or multiple “*and*”) such as “*MRI of joint of right lower extremity*,” it is even harder for the model to decide which word should be attached with the first one, which word should be attached with the second one, and the order of these two parts. Thus in this work, we also compared the performance of the LSTM-based model applied to the bags of words containing duplicate words and those which do not contain duplicate words. The result is shown in Table 7.5. The model achieved a better F1 score when applied to the bags of words without duplicates.

**Table 7.5:** Result of LSTM-based sequence prediction in terms of whether concept names contain duplicate words or not.

|                           | Without Duplicates | With Duplicates |
|---------------------------|--------------------|-----------------|
| Number of True Positives  | 4789               | 558             |
| Number of False Positives | 2224               | 533             |
| Number of False Negatives | 3101               | 312             |
| Precision                 | 68.29%             | 51.15%          |
| Recall                    | 60.70%             | 64.14%          |
| F1 Score                  | 64.27%             | 56.91%          |

### 7.3.2 Analysis of False Positives

We manually examined some of the false positive cases in Experiment II of binary classification for potential patterns that our model is not able to deal with. Meanwhile, we compared our sequence predictions for those false positive cases with the ground truth to explore possible causes for mislabeling. Two observed patterns are

listed as follows.

The first pattern is related to “*of*.” In SNOMED CT, “*B A*” and “*A of B*” are both acceptable names for certain cases. One of them is often defined as the fully specified name (FSN), while the other one is listed as its synonym. For instance, “*cyst of lung*” is an FSN, and “*lung cyst*” is its synonym. However, in some cases, only one of them is included (e.g., concept “*lung mass*” does not have a synonym “*mass of lung*”). Therefore, if a concept name falls into this pattern, our model sometimes cannot predict it correctly.

The second pattern involves two items (e.g., noun or noun phrase) that are connected by a preposition or conjunction, in which case our model sometimes cannot decide which one should be placed first. For example, for the concept “*naproxen sodium and sumatriptan*,” our predicted name was “*sumatriptan and naproxen sodium*.” Another example is that, for the concept “*fluoroscopy of left and right hip*,” our predicted name was “*fluoroscopy of right and left hip*.” In such cases, even the predicted names were valid, they were considered as false positives since they differ from the sequence of words provided in the ground truth. In other words, our evaluation was performed in a conservative way.

### 7.3.3 Beyond Naming Purpose

While analyzing false positive cases, we also noticed that this work could identify potential inconsistencies in the naming convention of concepts which can be considered as part of the quality assurance process for biomedical ontologies. For an existing name, we can use our trained model to check if it complies with the naming convention of SNOMED CT. For example, in the March 2018 US Edition of SNOMED CT, a new concept “*Lesion of bone right upper arm*” is added. Our model labeled it as wrong. We found that it is a synonym of “*Lesion of right upper arm bone*.” However, when it comes to another similar concept “*Lesion of left lower leg bone*,” it does not

have a synonym “*Lesion of bone left lower leg.*” Instead, it has “*Lesion of bone in left lower leg*” whose pattern does not appear as a synonym of “*Lesion of right upper arm bone.*”

Another example is “*Liver of normal size*” that has been added to SNOMED CT in the March 2018 Edition. Our model labeled it as wrong. We found that it is an FSN, and “*Normal sized liver*” is its synonym. However, in other similar concepts such as “*Normal sized tonsils*” and “*Normal sized ear canal,*” they are considered as FSNs, but they do not have any synonym that has the pattern “*xx of normal size.*” This indicates a naming inconsistency. A potential fix is that “*Normal sized liver*” should be the FSN, and the name “*Liver of normal size*” should become inactive. These two examples indicate that our model to some extent can reveal the inconsistency in SNOMED CT names.

## 7.4 Conclusion

In this chapter, we explore three deep learning-based approaches – simple neural network, LSTM, and CNN combined with LSTM, to predict concept names for new concepts given bags of words. Our experiments showed that the LSTM-based approach achieved the best performance with an F1 score of 63.41% for predicting names for newly added concepts in the March 2018 Edition of SNOMED CT and an F1 score of 73.95% for naming missing concepts identified by Cui et al.’s method in [22]. This indicates that the LSTM-based approach is effective in predicting concept names given bags of words. Further analysis of the false positive cases revealed that this work may also be leveraged for identifying potential inconsistencies within the concept names of SNOMED CT.

## CHAPTER 8. Preliminary Analysis of Cross-ontology Evaluation Based on Extrinsic Knowledge from UMLS

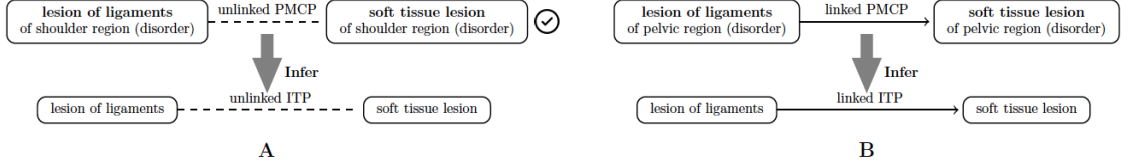
After quality issues were identified in biomedical ontologies, typically they will be manually reviewed by domain experts which is time-consuming and with a heavy workload.

In this dissertation, we explored different automated validation methods to relieve the manual burden. In the early-stage work of Chapter 4, to validate the uncovered missing hierarchical relations, we adopted the idea of Retrospective Ground Truth (RGT) proposed in [115] which leverages the difference between two versions of an ontology as the reference standard. More specifically, we used the newly added hierarchical relations in a newer version of NCI (19.07e inferred) compared with the 19.01d inferred version as the RGT, and then evaluated the potentially missing hierarchical relations suggested by our method against the RGT. In Chapter 6, we leveraged external ontologies in the UMLS to help validate the missing concepts identified by our FCA-based method. This kind of automatic validation method could show whether an auditing method is potentially effective.

In this chapter, we present a work in which cross-ontology verification based on extrinsic knowledge from UMLS is adopted to automatically validate missing hierarchical relations.

In this work, we first identify subtype inconsistencies within biomedical ontologies (Gene Ontology, NCI Thesaurus and SNOMED CT) by looking for identical linked and unlinked Inferred Term Pair (ITP) derived from linked and unlinked Partial Matching Concept Pair (PMCP). An example is shown in Figure 8.1. In SNOMED CT, concept “*Lesion of ligaments of pelvic region (disorder)*” is a subtype of “*Soft tissue lesion of pelvic region (disorder)*.” They two could form a lined PMCP and infer a linked ITP (“lesion of ligaments,” “soft tissue lesion”). Similarly, “*Lesion of*

*ligaments of shoulder region (disorder)*” is currently not a subtype of “*Soft tissue lesion of shoulder region (disorder)*,” they two form an unlined PMCP and infer an unlinked IPT (“lesion of ligaments,” “soft tissue lesion”). Obviously, subtype inconsistency occurs (either the former one should be unlinked or the later one should be linked).



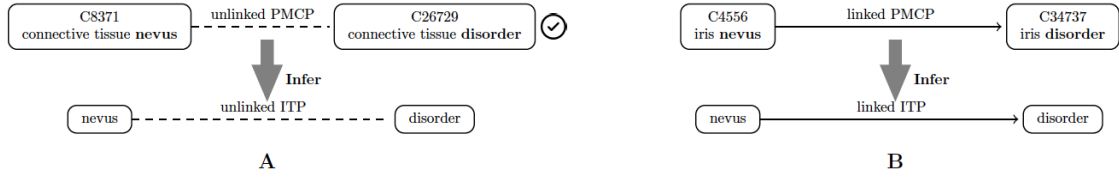
**Figure 8.1:** **A:** An unlinked PMCP with diff 3 in SNOMED CT and its unlinked IPT derived; **B:** A linked PMCP with diff 3 in SNOMED CT and its linked IPT derived. This example reveals a potentially **missing hierarchical relation** in **A**, that is, *lesion of ligaments of shoulder region (disorder)* IS-A *soft tissue lesion of shoulder region (disorder)*.

After detecting inconsistencies, besides manual review, we leverage external knowledge in UMLS (i.e., other ontologies in UMLS) to identify supporting evidence for detected potential subtype inconsistencies, which indicates the extent to which cross-ontology can help with validating whether a detected subtype inconsistency is a missing hierarchical relation. We performed such automated cross-ontology evaluation for Gene Ontology, NCI Thesaurus and SNOMED CT, respectively.

## 8.1 Method

Given an ontology, we perform a systematic check for each detected potential subtype inconsistency  $I$ . Assume that  $(u_1, u_2)$  is the unlinked PMCP involved in the inconsistency  $I$ . Then we map concepts  $u_1$  and  $u_2$  to the corresponding UMLS concepts  $m_1$  and  $m_2$ . If there exists a path  $p$  from  $m_1$  to  $m_2$  in UMLS such that  $p = m_1, m_{i_1}, m_{i_2}, \dots, m_{i_k}, m_2$  where  $m_1$  *is-a*  $m_{i_1}$ ,  $m_{i_1}$  *is-a*  $m_{i_2}$ ,  $\dots$ , and  $m_{i_k}$  *is-a*  $m_2$ , then we say that there is a piece of evidence in the UMLS supporting that  $u_1$  is a subtype of  $u_2$ . Note that the subtype relations along the path may be from different

ontologies. For instance, in Fig. 8.2, the path from “*connective tissues nevus*” (CUI: *C8371*) to “*connective tissue disorder*” (CUI: *C26729*) in UMLS was found through ontologies SNOMED CT and MEDCIN [116].



**Figure 8.2:** **A:** An unlinked PMCP with diff 3 in NCI Thesaurus and its unlinked ITP derived; **B:** A linked PMCP with diff 3 in NCI Thesaurus and its linked ITP derived. This example reveals a potentially **missing subtype relation** in **A**, that is, “*connective tissues nevus*” (CUI: *C8371*) IS-A “*connective tissue disorder*” (CUI: *C26729*).

## 8.2 Result

The UMLS-based evaluation identified supporting evidence for missing subtype relations involved in 26 detected inconsistencies in Gene Ontology, 306 in NCI Thesaurus, and 1,940 in SNOMED CT, respectively. Tables 8.1, 8.2 and 8.3 present ontologies in UMLS and their corresponding path contributions (PC) to identify supporting evidence for the detected potential subtype inconsistencies in Gene Ontology, NCI, and SNOMED CT, respectively. These tables contain the top 10 ontologies with the maximum path contributions. For Gene Ontology, Medical Subject Headings [117] contributed the most. For NCI Thesaurus, SNOMED CT contributed the most. For SNOMED CT, Read Thesaurus contributed the most.

In this preliminary work, cross-ontology evaluation only showed limited supporting evidence: 0.54% (=26/4841) for Gene Ontology, 11.43% (=306/2677) for NCI Thesaurus, and 3.61% (=1940/53782) for SNOMED CT. It would be interesting to further investigate methods to leverage other external knowledge such as biomedical literature to automatically identify supporting evidence for detected potential inconsistencies and reduce domain experts’ manual effort.

**Table 8.1:** Ontologies and corresponding Path Contributions (PC) for the UMLS-based evaluation of detected subtype inconsistencies in Gene Ontology.

| $n = 1$                                    |    | $n = 2$                          |    |
|--|----|----------------------------------|----|
| Ontology                                   | PC | Ontology                         | PC |
| Medical Subject Headings                   | 15 | NCIt                             | 3  |
| NCIt                                       | 11 | CRISP Thesaurus                  | 2  |
| Crisp Thesaurus                            | 11 | Alcohol and Other Drug Thesaurus | 1  |
| Alcohol and Other Drug Thesaurus           | 7  |                                  |    |
| Foundation Model of Anatomy Ontology       | 6  |                                  |    |
| LOINC                                      | 5  |                                  |    |
| Thesaurus of Psychological Index Terms     | 5  |                                  |    |
| SNOMED CT                                  | 5  |                                  |    |
| University of Washington Digital Anatomist | 5  |                                  |    |
| Read Thesaurus                             | 5  |                                  |    |

**Table 8.2:** Ontologies and corresponding Path Contributions (PC) for the UMLS-based evaluation of detected subtype inconsistencies in NCI Thesaurus.

| $n = 1$  |     | $n = 2$  |    |
|--|-----|--|----|
| Ontology   | PC  | Ontology   | PC |
| SNOMED CT  | 184 | SNOMED CT  | 48 |
| Read Thesaurus   | 86  | Read Thesaurus   | 31 |
| Medical Subject Headings   | 60  | MedDRA   | 15 |
| MEDCIN   | 57  | International Classification of Diseases Related Health Problems | 13 |
| MedDRA   | 44  | Medical Subject Headings   | 13 |
| National Drug File-Reference Terminology                             | 33  | MEDCIN   | 10 |
| CRISP Thesaurus  | 32  | National Drug File-Reference Terminology                         | 9  |
| Alcohol and Other Drug Thesaurus                                     | 24  | CRISP Thesaurus  | 9  |
| COSTART  | 22  | COSTART  | 9  |
| International Classification of Diseases and Related Health Problems | 17  | Human Phenotype Ontology   | 8  |

**Table 8.3:** Ontologies and corresponding Path Contributions (PC) for the UMLS-based evaluation of detected subtype inconsistencies in SNOMED CT.

| $n = 1$  |     | $n = 2$  |     | $n = 3$  |    | $n = 4$                                    |    |
|--|-----|--|-----|--|----|--|----|
| Ontology   | PC  | Ontology   | PC  | Ontology   | PC | Ontology                                   | PC |
| Read Thesaurus   | 954 | Read Thesaurus   | 283 | Read Thesaurus   | 49 | Read Thesaurus                             | 13 |
| MEDCIN   | 323 | MEDCIN   | 69  | MEDCIN   | 19 | Medical Subject Headings                   | 4  |
| NCIt   | 291 | Medical Subject Headings   | 63  | Foundational Model of Anatomy Ontology                               | 17 | NCIt                                       | 3  |
| Medical Subject Headings   | 257 | NCIt   | 61  | University of Washington Digital Anatomist                           | 17 | National Drug File - Reference Terminology | 3  |
| CRISP Thesaurus  | 164 | Alcohol and Other Drug Thesaurus                                     | 43  | NCIt   | 13 | CRISP Thesaurus                            | 2  |
| Alcohol and Other Drug Thesaurus                                     | 138 | CRISP Thesaurus  | 40  | International Classification of Diseases and Related Health Problems | 10 | University of Washington Digital Anatomist | 1  |
| National Drug File - Reference Terminology                           | 109 | National Drug File - Reference Terminology                           | 30  | Medical Subject Headings   | 10 | MEDCIN                                     | 1  |
| International Classification of Diseases and Related Health Problems | 85  | University of Washington Digital Anatomist                           | 29  | Human Phenotype Ontology   | 4  |  |    |
| Foundational Model of Anatomy Ontology                               | 84  | Foundational Model of Anatomy Ontology                               | 27  | National Drug File - Reference Terminology                           | 3  |  |    |
| MedDRA   | 77  | International Classification of Diseases and Related Health Problems | 21  | MedlinePlus Health Topics  | 2  |  |    |

## CHAPTER 9. Discussion, Conclusions and Future Directions

### 9.1 Discussion

In this dissertation, we introduced several automated and scalable approaches for ontology quality assurance. By “scalable,” we mean that the approach can be applied to the entire ontology and its performance (e.g., the precision of suggested missing hierarchical relations) remains compatible compared with being applied to a specific part or sub-hierarchies of the ontology. Existing studies are often confined to part of the ontology and do not scale to the entirety. For instance, previous study [25] regarded words appearing in the concept names as logical definitions, and compared lexical-derived hierarchy with the original hierarchy to reveal missing hierarchical relations. Their method achieved a decent precision while applying to two specific groups of concepts (concepts under “*Disorder of head (disorder)*” and “*Operative procedure on head (procedure)*”). However, if we apply the method to a different group of concepts (e.g., contains concepts from different knowledge branches) or to the entire SNOMED CT, many false positives will appear in the results. For example, related concepts (e.g., “*Erlotinib (substance)*” and “*Erlotinib hydrochloride (substance)*”), as well as un-related concepts (e.g., “*Acute pain (finding)*” and “*Acute sensitivity to pain (finding)*”) can both be incorrectly linked with hierarchical relation. When it comes to our approaches in auditing hierarchical relations, in Chapter 3 and Chapter 4, we leveraged a rich set of features (e.g., noun phrase and associative roles) to distinguish the semantic meanings of concepts so that those potential erroneous relations can be greatly relieved. Also, we applied our method to the entire ontology (either exhaustively or by first recognizing problematic sub-structures), and then randomly selected samples for manual review which ensured that the precision of the sampled missing hierarchical relations could reflect the effectiveness of our approaches in auditing the entire ontology.



Next we briefly discuss the time complexity of our quality assurance approaches. An ontology can be considered as a directed graph  $G(V, E)$ , where  $V$  is a set of nodes representing the concepts and  $E$  is a set of edges representing relations among concepts (e.g., hierarchical relations). Regarding the auditing approaches introduced in Chapter 3 and Chapter 4, their time complexity is  $O(|V|^2)$  if we pre-compute the lexical features or logical definitions of concepts (i.e., retrieve information from pre-constructed hash tables which support constant time look up operation) and then perform pairwise comparison among  $|V|$  concepts. In Chapter 5, to identify potentially missing hierarchical relations, we performed name transformation for each concept – noun chunks in the concept name were replaced by more general terms to generate potential supertype concepts. Therefore, the time complexity of this auditing approach can be computed by summing up the transformation time  $T_i$  for each concept  $i \in V$ . Here  $T_i = \prod_{j \in NP_i} |R_{ij}|$ , where  $NP_i$  denotes the set of noun chunks to be replaced in  $i$ 's concept name and  $R_{ij}$  denotes the set of replacement candidates for noun chunk  $j \in NP_i$ . Let  $n$  be the maximum number of noun chunks to be replaced in concept names and  $r$  be the maximum number of replacement candidates for noun chunks respectively, then  $\sum_{i \in V} T_i$  is bounded by  $r^n \cdot |V|$ . Regarding the missing hierarchical relations identified in Chapter 5, the subtype concepts on which we perform the transformation are all with no larger than three noun chunks in their concept names (i.e.,  $n = 3$ ); although the maximum number of replacement candidates for noun chunks is 40 (i.e.,  $r = 40$ ), the average number is less than 3. Therefore, the running time of the transformation-based approach in our experiment is close to the best case  $O(|V|)$ .

## 9.2 Conclusions

Biomedical ontologies play vital roles in downstream biomedical applications. Biomedical ontologies are constantly evolving and thus usually incomplete. However, the lack

of completeness may be unacceptable for applications in areas such as healthcare and defense, where missing answers can adversely affect the application’s functionality. Due to the sheer size and complexity of biomedical ontologies, manually auditing the completeness is of poor efficiency. Therefore, automated frameworks that can uncover or validate the incompleteness issues in biomedical ontologies are highly desirable.

This dissertation introduces scalable approaches for identifying and validating potential incompleteness issues (i.e., missing hierarchical relations and missing concepts) in biomedical ontologies using a combination of (1) logical definitions and lexical features of concepts; (2) mathematical underpinning with Formal Concept Analysis; and (3) extrinsic knowledge sources.

Chapter 3 introduces a lexical-based approach to automatically detect potentially missing hierarchical relations. We model each concept with an enriched set of lexical features, by leveraging words and noun phrases in the name of the concept itself and the concept’s ancestors. Then we perform subset inclusion checking on enriched lexical feature sets to suggest potentially missing hierarchical relations between concepts. We applied our approach to the September 2017 release of SNOMED CT (US edition) which suggested a total of 38,615 potentially missing hierarchical relations. For evaluation, a domain expert reviewed a random sample of 100 missing hierarchical relations selected from the “*Clinical finding*” sub-hierarchy, and confirmed 90 are valid, indicating that our method achieved a precision of 90% in detecting missing hierarchical relations. Additional review of invalid suggestions further revealed incorrect existing hierarchical relations. Our results demonstrated that systematic analysis of the enriched lexical features of concepts is an effective approach to identify potentially missing hierarchical relations in the SNOMED CT.

Chapter 4 presents a framework utilizing lexical features and role definitions of concepts to identify missing hierarchical relations in non-lattice subgraphs. Regarding the method, we first compute all the non-lattice subgraphs (i.e., areas that are

likely to contain quality issues). Then, we model each concept using its associative roles, words and roots of noun chunks within its concept name and its ancestor’s names. At last, we perform subsumption testing for candidate concept pairs in the non-lattice subgraphs to automatically detect potentially missing hierarchical relations. We applied our approach to the 19.08d version of the NCI Thesaurus. A total of 55 potentially missing hierarchical relations were identified by our approach. Domain experts confirmed 29 out of 55 as valid and incorporated them in the newer versions of the NCI Thesaurus. 7 out of 55 further revealed incorrect existing hierarchical relations in the NCI Thesaurus. The results showed that leveraging both lexical features and role definitions benefits semantic modeling of concepts as well as incompleteness detection.

Chapter 5 introduces a novel transformation-based auditing method that leverages the UMLS knowledge to systematically identify missing hierarchical relations in its source ontologies. Given a concept name in the UMLS, we first identify its base and secondary noun chunks. For each identified noun chunk, we generate replacement candidates that are more general than the noun chunk. Then we replace the noun chunks with their replacement candidates to generate new potential concept names which may serve as supertypes of the original concept. If a newly generated name is an existing concept name in the same source ontology with the original concept, then a potentially missing hierarchical relation between the original and the new concept is identified. Applying our transformation-based method to English-language concept names in the UMLS (2019AB release), a total of 39,359 potentially missing hierarchical relations were detected in 13 source ontologies. Domain experts evaluated a random sample of 200 potentially missing hierarchical relations identified in the SNOMED CT (US edition), and 100 in the Gene Ontology. 173 out of 200 and 63 out of 100 potentially missing IS-A relations were confirmed by domain experts, indicating our method achieved a precision of 86.5% and 63% for the SNOMED CT

and Gene Ontology, respectively.

Chapter 6 introduces a lexical method based on Formal Concept Analysis (FCA) to identify potentially missing concepts in a given ontology by leveraging its intrinsic knowledge – concept names. Lexical features (i.e., words appearing in the concept names) are considered as FCA attributes while generating formal context. Applying multistage intersection on FCA attributes identifies newly formalized bags of words (i.e., FCA formal concepts) that represent missing concepts, which may be further validated through external knowledge. We applied our method to the *Disease or Disorder* sub-hierarchy in the 19.08d version of the NCI Thesaurus and identified a total of 8,983 potentially missing concepts. The preliminary evaluation via external validation using UMLS showed encouraging evidence for the effectiveness of our method.

Chapter 7 shows deep learning-based approaches, given bags of words, to automatically predict concept names that comply with the naming convention of SNOMED CT. These deep learning models are simple neural network, Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN) combined with LSTM. Our experiments showed that LSTM-based approach achieved the best performance: a precision of 65.98%, a recall of 61.04%, and an F1 score of 63.41% for predicting concept names for newly added concepts in the March 2018 Edition of SNOMED CT. It also achieved a precision of 74.58%, a recall of 73.33%, and an F1 score of 73.95% for naming missing concepts identified by our previous work. Further examination of results revealed inconsistencies within SNOMED CT which may be leveraged for quality assurance purposes.

In Chapter 8, we discuss the possibility to use automatic validating methods based on Retrospective Ground Truth (RGT) and extrinsic knowledge from the UMLS to relieve the heavy work of manual review. We also perform a preliminary study on the extent to which external knowledge in the UMLS can provide supporting evidence for

validating the detected missing hierarchical relations.

## **9.3 Future directions**

### **9.3.1 Repair Missing Hierarchical Relations**

Given a set of missing hierarchical relations, the simplest remediation measure is to just add them into the ontology hierarchy. However, since most auditing methods rely on the inferred definitions of concepts, the missing hierarchical relations they detected also pertain to the inferred hierarchy which is obtained by reasoners based on the stated logical definitions. Therefore, it is more meaningful to investigate why the missing hierarchical relations are not derived by the DL reasoners. Recently, we developed a method based on semantic-related group pairs to reveal the causes of missing hierarchical relations, as well as locate quality issues in the stated logical definitions. After repairing those quality issues, the missing hierarchical relations will also become derivable by the reasoners and thus fixed. Compared with purely adding the missing hierarchical relations, this method could better improve the quality of biomedical ontologies. We plan to apply our method to the missing hierarchical relations detected in our previous work and provide our result to the domain experts so that a comprehensive evaluation of our method could be provided in the near future.

### **9.3.2 Improved Formal Concept Analysis**

Regarding the lexical-based FCA method proposed in Chapter 6, since we consider each word in the concept name as individual feature, the results from multistage intersection (i.e., formalizing concepts) only includes bags of words. Although we have found some supporting evidence (i.e., matching concept names) in the UMLS, large portion of revealed missing concepts are still waiting for evaluation to prove the effectiveness of our method. However, it is inconvenient for domain experts to evaluate

bags of words. To name the concepts, we could leverage the work in Chapter 7 to predict concept names given bags of words. In the other way, we could re-define the attributes used in FCA and the process of multistage intersection so that the attributes we get for the new concepts are meaningful sequences (i.e., concept names). Currently, formal context adopts sets of attributes (e.g., sets of words) as input and the intersection is performed between sets. In the future, we plan to consider directly using the whole concept names (i.e., sequences) as the attributes. Then the FCA results rely on how to formalize new concepts from the “intersection” (which previously is defined as intersection between sets) between concept names. An intuitive idea is to find the common sub-strings between two concept names. On the other hand, we have studied different layouts of a concept name (e.g., breaking the concept name as a combination of noun phrases and words used in Chapter 5 or sequence representation based on sub-term and pos-tagging in [96]). We could utilize these variants and re-define the concept-forming operator in traditional FCA to generate either more general or more detailed concept names.

### 9.3.3 Deep learning approaches

In Chapter 7, we adopted several basic deep learning models to predict concept names. However, the sequence prediction method is based on classifying all possible combinations of words, which is computationally challenging when the concept name is long. Therefore, we plan to build a more powerful generative neural network model to generate potential concept names based on descriptive input.

In Chapter 4, we have explored how to harmonize the logical definitions with lexical features to represent the semantic meanings of concepts. The hybrid model could be adapted to embeddings of concepts and thus potentially be a determinant in how to name the concepts. We plan to build a sequence-to-sequence neural network which takes our hybrid semantic model as input and generates corresponding concept

names as outputs. In this case, we no longer need to rely on the bags of words and test all the possible combinations. In addition, a large portion of non-lattice subgraphs in Chapter 4 is not covered by rule-based auditing methods. Therefore, we also plan to leverage machine learning techniques to uncover additional missing hierarchical relations for those unexplored non-lattice subgraphs.

#### **9.3.4 Automatic validation method**

In this dissertation, to automatically validate the detected incompleteness issues, we mainly utilized the extrinsic knowledge from the UMLS. Another rich source that could be leveraged is biomedical literature. For example, MEDLINE/PubMed comprises more than 27 million records representing articles in the biomedical literature, which documents article titles, abstract and controlled vocabulary search terms [118]. In the future, we plan to develop entity-relations recognition approaches to extract evidence from biomedical literature to support validation of missing hierarchical relations.

## REFERENCES

- [1] Olivier Bodenreider. Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearbook of medical informatics*, 17(01):67–79, 2008.
- [2] Robert Hoehndorf, Paul N Schofield, and Georgios V Gkoutos. The role of ontologies in biological and biomedical research: a functional perspective. *Briefings in bioinformatics*, 16(6):1069–1080, 2015.
- [3] National Cancer Institute Thesaurus. <https://ncit.nci.nih.gov/>. [Accessed 10-January-2021].
- [4] Sherri De Coronado, Margaret W Haber, Nicholas Sioutos, Mark S Tuttle, Lawrence W Wright, et al. Nci thesaurus: using science-based terminology to integrate cancer research results. In *Medinfo*, pages 33–37, 2004.
- [5] Gilberto Fragoso, Sherri de Coronado, Margaret Haber, Frank Hartel, and Larry Wright. Overview and utilization of the nci thesaurus. *Comparative and functional genomics*, 5(8):648–654, 2004.
- [6] Nicholas Sioutos, Sherri de Coronado, Margaret W Haber, Frank W Hartel, Wen-Ling Shaiu, and Lawrence W Wright. Nci thesaurus: a semantic model integrating cancer-related clinical and molecular information. *Journal of biomedical informatics*, 40(1):30–43, 2007.
- [7] SNOMED CT. <https://www.snomed.org/>. [Accessed 10-January-2021].
- [8] Dennis Lee, Nicolette de Keizer, Francis Lau, and Ronald Cornet. Literature review of snomed ct use. *Journal of the American Medical Informatics Association*, 21(e1):e11–e19, 2013.
- [9] Rainer Winnenburg and Olivier Bodenreider. Metrics for assessing the quality of value sets in clinical quality measures. In *AMIA Annual Symposium Proceedings*, volume 2013, page 1497. American Medical Informatics Association, 2013.
- [10] Licong Cui, Shiqiang Tao, and Guo-Qiang Zhang. Biomedical ontology quality assurance using a big data approach. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 10(4):41, 2016.
- [11] B Cuenca Grau, Boris Motik, Giorgos Stoilos, and Ian Horrocks. Completeness guarantees for incomplete ontology reasoners: Theory and practice. *Journal of Artificial Intelligence Research*, 43:419–476, 2012.
- [12] SNOMED International Release Management Home. <https://confluence.ihitsdotools.org/display/RMT/>. [Accessed 10-January-2021].



- [13] Overview of NCI Thesaurus (NCIt). <https://wiki.nci.nih.gov/pages/viewpage.action?pageId=7472532>. [Accessed 10-January-2021].
- [14] Patrick Lambrix, Fang Wei-Kleiner, and Zlatan Dragisic. Completing the is-a structure in light-weight ontologies. *Journal of biomedical semantics*, 6(1):12, 2015.
- [15] Licong Cui, Remo Mueller, Satya Sahoo, and Guo-Qiang Zhang. Querying complex federated clinical data using ontological mapping and subsumption reasoning. In *2013 IEEE International Conference on Healthcare Informatics*, pages 351–360. IEEE, 2013.
- [16] Zhe He, James Geller, and Yan Chen. A comparative analysis of the density of the snomed ct conceptual content for semantic harmonization. *Artificial intelligence in medicine*, 64(1):29–40, 2015.
- [17] Zhe He, Yan Chen, Sherri de Coronado, Katrina Piskorski, and James Geller. Topological-pattern-based recommendation of umls concepts for national cancer institute thesaurus. In *AMIA Annual Symposium Proceedings*, volume 2016, page 618. American Medical Informatics Association, 2016.
- [18] Guoqian Jiang and Christopher G Chute. Auditing the semantic completeness of snomed ct using formal concept analysis. *Journal of the American Medical Informatics Association*, 16(1):89–102, 2009.
- [19] Zhu Wei, Cui Licong, and Zhang Guo-Qiang. Spark-mca: Large-scale, exhaustive formal concept analysis for evaluating the semantic completeness of snomed ct. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1931. American Medical Informatics Association, 2017.
- [20] Praveen Chandar, Anil Yaman, Julia Hoxha, Zhe He, and Chunhua Weng. Similarity-based recommendation of new concepts to a terminology. In *AMIA Annual Symposium Proceedings*, volume 2015, page 386. American Medical Informatics Association, 2015.
- [21] Jiajie Peng, Tao Wang, Jixuan Wang, Yadong Wang, and Jin Chen. Extending gene ontology with gene association networks. *Bioinformatics*, 32(8):1185–1194, 2016.
- [22] Licong Cui, Wei Zhu, Shiqiang Tao, James T Case, Olivier Bodenreider, and Guo-Qiang Zhang. Mining non-lattice subgraphs for detecting missing hierarchical relations and concepts in snomed ct. *Journal of the American Medical Informatics Association*, 24(4):788–798, 2017.
- [23] Rashmie Abeysinghe, Michael A Brooks, Jeffery Talbert, and Cui Licong. Quality assurance of nci thesaurus by mining structural-lexical patterns. In *AMIA Annual Symposium Proceedings*, volume 2017, page 364. American Medical Informatics Association, 2017.

- [24] Yan Chen, Huanying Helen Gu, Yehoshua Perl, and James Geller. Structural group-based auditing of missing hierarchical relationships in umls. *Journal of biomedical informatics*, 42(3):452–467, 2009.
- [25] Olivier Bodenreider. Identifying missing hierarchical relations in snomed ct from logical definitions based on the lexical features of concept names. *ICBO/BioCreative*, 2016, 2016.
- [26] Licong Cui, Olivier Bodenreider, Jay Shi, and Guo-Qiang Zhang. Auditing snomed ct hierarchical relations based on lexical features of concepts in non-lattice subgraphs. *Journal of biomedical informatics*, 78:177–184, 2018.
- [27] Rashmie Abeysinghe, Eugene W Hinderer, Hunter NB Moseley, and Licong Cui. Auditing subtype inconsistencies among gene ontology concepts. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1242–1245. IEEE, 2017.
- [28] Rashmie Abeysinghe, Fengbo Zheng, Eugene W Hinderer, Hunter NB Moseley, and Licong Cui. A lexical approach to identifying subtype inconsistencies in biomedical terminologies. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1982–1989. IEEE, 2018.
- [29] Manuel Quesada-Martínez, Jesualdo Tomás Fernández-Breis, and Daniel Karlsson. Suggesting missing relations in biomedical ontologies based on lexical regularities. In *MIE*, pages 384–388, 2016.
- [30] Vipina K Keloth, Zhe He, Yan Chen, and James Geller. Leveraging horizontal density differences between ontologies to identify missing child concepts: A proof of concept. In *AMIA Annual Symposium Proceedings*, volume 2018, page 644. American Medical Informatics Association, 2018.
- [31] Hao Liu, Ling Zheng, Yehoshua Perl, James Geller, and Gai Elhanan. Can a convolutional neural network support auditing of nci thesaurus neoplasm concepts? In *ICBO*, 2018.
- [32] Qi Sun, Guo-Qiang Zhang, Wei Zhu, and Licong Cui. Validating auto-suggested changes for snomed ct in non-lattice subgraphs using relational machine learning. *Studies in health technology and informatics*, 2019.
- [33] Christopher Ochs, Zhe He, Ling Zheng, James Geller, Yehoshua Perl, George Hripcsak, and Mark A Musen. Utilizing a structural meta-ontology for family-based quality assurance of the bioportal ontologies. *Journal of biomedical informatics*, 61:63–76, 2016.
- [34] H Gu, Y Chen, Z He, M Halper, and L Chen. Quality assurance of umls semantic type assignments using snomed ct hierarchies. *Methods of information in medicine*, 55(02):158–165, 2016.

- [35] Christopher Ochs, James Geller, Yehoshua Perl, Yan Chen, Ankur Agrawal, James T Case, and George Hripcsak. A tribal abstraction network for snomed ct target hierarchies without attribute relationships. *Journal of the American Medical Informatics Association*, 22(3):628–639, 2014.
- [36] Christopher Ochs, James Geller, Yehoshua Perl, Yan Chen, Junchuan Xu, Hua Min, James T Case, and Zhi Wei. Scalable quality assurance for large snomed ct hierarchies using subject-based subtaxonomies. *Journal of the American Medical Informatics Association*, 22(3):507–518, 2015.
- [37] F Zheng, J Shi, and L Cui. A lexical-based approach for exhaustive detection of missing hierarchical is-a relations in snomed ct. In *AMIA 2020 Annual Symposium Proceedings*. American Medical Informatics Association, [In Press].
- [38] Fengbo Zheng, Rashmie Abeysinghe, Nicholas Sioutos, Lori Whiteman, Lyubov Remennik, and Licong Cui. Detecting missing is-a relations in the nci thesaurus using an enhanced hybrid approach. *BMC Medical Informatics and Decision Making*, 20(10):1–11, 2020.
- [39] Fengbo Zheng, Rashmie Abeysinghe, and Licong Cui. A hybrid method to detect missing hierarchical relations in nci thesaurus. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1948–1953. IEEE, 2019.
- [40] Fengbo Zheng, Jay Shi, Yuntao Yang, W Jim Zheng, and Licong Cui. A transformation-based method for auditing the is-a hierarchy of biomedical terminologies in the unified medical language system. *Journal of the American Medical Informatics Association*, 27(10):1568–1575, 2020.
- [41] Fengbo Zheng and Licong Cui. A lexical-based formal concept analysis method to identify missing concepts in the nci thesaurus. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1757–1760. IEEE, 2020.
- [42] Fengbo Zheng and Licong Cui. Exploring deep learning-based approaches for predicting concept names in snomed ct. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 808–813. IEEE, 2018.
- [43] Rashmie Abeysinghe, Michael A Brooks, and Licong Cui. Leveraging non-lattice subgraphs to audit hierarchical relations in nci thesaurus. In *AMIA Annual Symposium Proceedings*, volume 2019, page 982. American Medical Informatics Association, 2019.
- [44] Dieter Fensel. *Ontologies: a silver bullet for knowledge management and electronic commerce*. Springer, 2011.
- [45] Natalya F Noy, Nigam H Shah, Patricia L Whetzel, Benjamin Dai, Michael Dorf, Nicholas Griffith, Clement Jonquet, Daniel L Rubin, Margaret-Anne

- Storey, Christopher G Chute, et al. Biportal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, 37(suppl\_2):W170–W173, 2009.
- [46] BioPortal. <https://bioportal.bioontology.org/>. [Accessed 10-January-2021].
  - [47] Manuel Salvadores, Paul R Alexander, Mark A Musen, and Natalya F Noy. Biportal as a dataset of linked biomedical ontologies and terminologies in rdf. *Semantic web*, 4(3):277–284, 2013.
  - [48] Thomas R Gruber. Toward principles for the design of ontologies used for knowledge sharing? *International journal of human-computer studies*, 43(5-6):907–928, 1995.
  - [49] Yevgeny Kazakov, Markus Krötzsch, and Frantisek Simancik. Elk reasoner: Architecture and evaluation. In *ORE*, 2012.
  - [50] Michael J Lawley and Cyril Bousquet. Fast classification in protégé: Snorocket as an owl 2 el reasoner. In *Proc. 6th Australasian Ontology Workshop (IAOA’10). Conferences in Research and Practice in Information Technology*, volume 122, pages 45–49, 2010.
  - [51] SNOMED CT Starter Guide. <https://confluence.ihtsdotools.org/display/DOCSTART/SNOMED+CT+Starter+Guide>. [Accessed 10-January-2021].
  - [52] SNOMED CT Browser. <https://browser.ihtsdotools.org/>. [Accessed 10-January-2021].
  - [53] SNOMED CT Diagram Guideline. <https://confluence.ihtsdotools.org/display/DOCDIAG/Diagramming+Guideline>. [Accessed 10-January-2021].
  - [54] Frank W Hartel, Sherri de Coronado, Robert Dionne, Gilberto Fragoso, and Jennifer Golbeck. Modeling a description logic vocabulary for cancer research. *Journal of biomedical informatics*, 38(2):114–129, 2005.
  - [55] Melissa A Haendel, Julie A McMurry, Rose Relevo, Christopher J Mungall, Peter N Robinson, and Christopher G Chute. A census of disease ontologies. *Annual Review of Biomedical Data Science*, 1:305–331, 2018.
  - [56] Vocabulary for Cancer Research. <https://datascience.cancer.gov/resources/cancer-vocabulary>. [Accessed 10-January-2021].
  - [57] Gene Ontology General Overview. <http://geneontology.org/docs/faq/>. [Accessed 10-January-2021].
  - [58] Gene Ontology Consortium. The gene ontology project in 2008. *Nucleic acids research*, 36(suppl\_1):D440–D444, 2008.

- [59] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.
- [60] Christophe Dessimoz and Nives Škunca. *The Gene Ontology Handbook*. Humana Press New York, NY, USA:, 2017.
- [61] The Gene Ontology Resource. <http://geneontology.org/>. [Accessed 10-January-2021].
- [62] Dmitry I Ignatov. Introduction to formal concept analysis and its applications in information retrieval and related fields. In *Russian Summer School in Information Retrieval*, pages 42–141. Springer, 2014.
- [63] Bernhard Ganter and Rudolf Wille. *Formal Concept Analysis: mathematical foundations*. Springer Science & Business Media, 2012.
- [64] Guo-Qiang Zhang and Olivier Bodenreider. Large-scale, exhaustive lattice-based structural auditing of snomed ct. In *AMIA annual symposium proceedings*, volume 2010, page 922. American Medical Informatics Association, 2010.
- [65] Li Deng, Dong Yu, et al. Deep learning: methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4):197–387, 2014.
- [66] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [67] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232, 2017.
- [68] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005.
- [69] Canlin Zhang, Daniel Biš, Xiuwen Liu, and Zhe He. Biomedical word sense disambiguation with bidirectional long short-term memory and attention-based neural networks. *BMC bioinformatics*, 20(16):1–15, 2019.
- [70] Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis Lau. A c-lstm neural network for text classification. *arXiv preprint arXiv:1511.08630*, 2015.
- [71] Christopher G Chute, Yiming Yang, and DA Evans. Latent semantic indexing of medical diagnoses using umls semantic structures. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 185. American Medical Informatics Association, 1991.

- [72] Prakash Nadkarni, Roland Chen, and Cynthia Brandt. Umls concept indexing for production databases: a feasibility study. *Journal of the American Medical Informatics Association*, 8(1):80–91, 2001.
- [73] William Hersh, Susan Price, and Larry Donohoe. Assessing thesaurus-based query expansion using the umls metathesaurus. In *Proceedings of the AMIA Symposium*, page 344. American Medical Informatics Association, 2000.
- [74] Kun Lu and Xiangming Mu. Query expansion using umls tools for health information retrieval. *Proceedings of the American Society for Information Science and Technology*, 46(1):1–16, 2009.
- [75] David Martinez, Arantxa Otegi, Aitor Soroa, and Eneko Agirre. Improving search over electronic health records using umls-based query expansion through random walks. *Journal of biomedical informatics*, 51:100–106, 2014.
- [76] Alexa T McCray, Alan R Aronson, Allen C Browne, Thomas C Rindfleisch, Amir Razi, and Suresh Srinivasan. Umls knowledge for biomedical language processing. *Bulletin of the Medical Library Association*, 81(2):184, 1993.
- [77] Alan R Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.
- [78] Long Chen, Yu Gu, Xin Ji, Chao Lou, Zhiyong Sun, Haodan Li, Yuan Gao, and Yang Huang. Clinical trial cohort selection based on multi-level rule-based natural language processing system. *Journal of the American Medical Informatics Association*, 26(11):1218–1226, 2019.
- [79] Liang Yao, Chengsheng Mao, and Yuan Luo. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC medical informatics and decision making*, 19(3):71, 2019.
- [80] Ramon Maldonado, Meliha Yetisgen, and Sanda M Harabagiu. Adversarial learning of knowledge embeddings for the unified medical language system. *AMIA Summits on Translational Science Proceedings*, 2019:543, 2019.
- [81] Tomasz Adamusiak, Naoki Shimoyama, and Mary Shimoyama. Next generation phenotyping using the unified medical language system. *JMIR medical informatics*, 2(1):e5, 2014.
- [82] Soumeiya L Achour, Michel Dojat, Claire Rieux, Philippe Bierling, and Eric Lepage. A umls-based knowledge acquisition tool for rule-based clinical decision support system development. *Journal of the American Medical Informatics Association*, 8(4):351–360, 2001.
- [83] Pei-ju Lee, Yen-Hsien Lee, Yihuang Kang, and Ching-Ping Chao. A medical decision support system using text mining to compare electronic medical records.

- In *International Conference on Human-Computer Interaction*, pages 199–208. Springer, 2019.
- [84] Betsy L Humphreys and Donald AB Lindberg. Building the unified medical language system. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 475. American Medical Informatics Association, 1989.
  - [85] Donald AB Lindberg, Betsy L Humphreys, and Alexa T McCray. The unified medical language system. *Methods of information in medicine*, 32(4):281, 1993.
  - [86] Betsy L Humphreys, Donald AB Lindberg, Harold M Schoolman, and G Octo Barnett. The unified medical language system: an informatics research collaboration. *Journal of the American Medical Informatics Association*, 5(1):1–11, 1998.
  - [87] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl\_1):D267–D270, 2004.
  - [88] UMLS Reference Manual. <https://www.ncbi.nlm.nih.gov/books/NBK9676/>. [Accessed 10-January-2021].
  - [89] Guo-Qiang Zhang and Olivier Bodenreider. Using sparql to test for lattices: application to quality assurance in biomedical ontologies. In *International Semantic Web Conference*, pages 273–288. Springer, 2010.
  - [90] Rashmi Burse, G. Mcardle, and M. Bertolotto. Stop-word based contextual auditing to identify inconsistencies in snomed. In *SWH@ISWC*, 2020.
  - [91] Karin Verspoor, Daniel Dvorkin, K Bretonnel Cohen, and Lawrence Hunter. Ontology quality assurance through analysis of term transformations. *Bioinformatics*, 25(12):i77–i84, 2009.
  - [92] Michael Halper, Huanying Gu, Yehoshua Perl, and Christopher Ochs. Abstraction networks for terminologies: supporting management of “big knowledge”. *Artificial intelligence in medicine*, 64(1):1–16, 2015.
  - [93] Duo Wei and Olivier Bodenreider. Using the abstraction network in complement to description logics for quality assurance in biomedical terminologies-a case study in snomed ct. *Studies in health technology and informatics*, 160(0 2):1070, 2010.
  - [94] Christopher Ochs, Yehoshua Perl, and James Geller. Blusno: A system for orientation, visualization, and quality assurance of snomed ct using abstraction networks. In *ICBO*, pages 128–129, 2013.
  - [95] Hua Min, Yehoshua Perl, Yan Chen, Michael Halper, James Geller, and Yue Wang. Auditing as part of the terminology design life cycle. *Journal of the American Medical Informatics Association*, 13(6):676–690, 2006.

- [96] Rashmie Abeysinghe, Eugene W Hinderer III, Hunter NB Moseley, and Licong Cui. Ssif: Subsumption-based sub-term inference framework to audit gene ontology. *Bioinformatics*, 36(10):3207–3214, 2020.
- [97] Hao Liu, Yehoshua Perl, and James Geller. Transfer learning from bert to support insertion of new concepts into snomed ct. In *AMIA Annual Symposium Proceedings*, volume 2019, page 1129. American Medical Informatics Association, 2019.
- [98] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
- [99] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.
- [100] Guo-Qiang Zhang, Guangming Xing, and Licong Cui. An efficient, large-scale, non-lattice-detection algorithm for exhaustive structural auditing of biomedical ontologies. *Journal of biomedical informatics*, 80:106–119, 2018.
- [101] SpaCy: Industrial-Strength Natural Language Processing. <https://spacy.io/>. [Accessed 10-January-2021].
- [102] Licong Cui, Rashmie Abeysinghe, Fengbo Zheng, Shiqiang Tao, Ningzhou Zeng, Isaac Hands, Eric B Durbin, Lori Whiteman, Lyubov Remennik, Nicholas Sioutos, et al. Enhancing the quality of hierarchic relations in the national cancer institute thesaurus to enable faceted query of cancer registry data. *JCO clinical cancer informatics*, 4:392–398, 2020.
- [103] Spacy Linguistic Features. <https://spacy.io/usage/linguistic-features>. [Accessed 10-January-2021].
- [104] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.
- [105] LuiNorm. <https://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lvg/2021/docs/userDoc/tools/luiNorm.html>. [Accessed 10-January-2021].
- [106] Adam D Troy, Guo-Qiang Zhang, and Ye Tian. Faster concept analysis. In *International Conference on Conceptual Structures*, pages 206–219. Springer, 2007.
- [107] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.



- [108] Radim Rehurek and Petr Sojka. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer, 2010.
- [109] Christopher D. Manning Jeffrey Pennington, Richard Socher. Keras. <https://keras.io>, 2015.
- [110] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [111] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*, 2014.
- [112] Dan Ciregan, Ueli Meier, and Jürgen Schmidhuber. Multi-column deep neural networks for image classification. In *Computer vision and pattern recognition (CVPR), 2012 IEEE conference on*, pages 3642–3649. IEEE, 2012.
- [113] G David Forney. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973.
- [114] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.
- [115] Guo-Qiang Zhang, Huang Yan, and Licong Cui. Can snomed ct changes be used as a surrogate standard for evaluating the performance of its auditing methods? In *AMIA Annual Symposium Proceedings*, volume 2017, pages 1903–1912. American Medical Informatics Association, 2017.
- [116] MEDCIN. <https://medicomp.com/>. [Accessed 10-January-2021].
- [117] Medical Subject Headings. <https://www.nlm.nih.gov/mesh/meshhome.html>. [Accessed 10-January-2021].
- [118] PubMed Online Training. <https://www.nlm.nih.gov/bsd/disted/pubmedtutorial/cover.html>. [Accessed 10-January-2021].

## Vita

### Personal Information

- Name: Fengbo Zheng
- Place of birth: Tianjin, China

### Education

- Master of Science in Computer Engineering, University of Florida, Gainesville, FL, USA, May 2016
- Bachelor of Engineering in Software Engineering, Wuhan University, Hubei, China, June 2014

### Professional Experience

- Scientific Programmer, University of Texas Health Science Center at Houston Houston, TX, USA. Oct 2020 - Jun 2021
- Visiting Student Trainee, University of Texas Health Science Center at Houston Houston, TX, USA. Aug 2019 - May 2020
- Research Assistant, University of Kentucky Lexington, KY, USA. Jan 2018 - Aug 2019
- Teaching Assistant, University of Kentucky Lexington, KY, USA. Aug 2016 - Dec 2017

### Scholastic and Professional Awards

- Department of Computer Science, University of Kentucky Student Travel Award. BIBM 2019, San Diego, CA, USA
- Department of Computer Science, University of Kentucky Student Travel Award. BIBM 2018, Madrid, Spain
- University of Florida Achievement Award Scholarship. 2014

### Publications

1. Fengbo Zheng, Jay Shi, Licong Cui. A Lexical-based Approach for Exhaustive Detection of Missing Hierarchical IS-A Relations in SNOMED CT. In *AMIA Annual Symposium Proceedings 2020*, page 1392-1401, 2020.
2. Fengbo Zheng, Licong Cui. A Lexical-based Formal Concept Analysis Method to Identify Missing Concepts in the NCI Thesaurus. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, page 1757-1760, 2020.

3. Fengbo Zheng, Jay Shi, Yuntao Yang, W. Jim Zheng, Licong Cui. A Transformation-based Method for Auditing the IS-A Hierarchy of Biomedical Terminologies in the Unified Medical Language System. In *Journal of the American Medical Informatics Association*, 27(10): 1568-1575, 2020.
4. Fengbo Zheng, Rashmie Abeysinghe, Nicholas Sioutos, Lori Whiteman, Lyubov Remennik, Licong Cui. Detecting Missing IS-A Relations in the NCI Thesaurus Using an Enhanced Hybrid Approach. *BMC medical informatics and decision making*, 20-S(10):273, 2020.
5. Licong Cui, Rashmie Abeysinghe, Fengbo Zheng, Shiqiang Tao, Ningzhou Zeng, Isaac Hands, Eric B Durbin, Lori Whiteman, Lyubov Remennik, Nicholas Sioutos, Guo-Qiang Zhang. Enhancing the Quality of Hierarchic Relations in the National Cancer Institute Thesaurus to Enable Faceted Query of Cancer Registry Data. *JCO Clinical Cancer Informatics*, 4: 392-398, 2020.
6. Fengbo Zheng, Rashmie Abeysinghe, and Licong Cui. A Hybrid Method to Detect Missing Hierarchical Relations in NCI Thesaurus. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, page 1948-1953, 2019.
7. Jing Liu, Rashmie Abeysinghe, Fengbo Zheng, Licong Cui. Pattern-based Extraction of Disease Drug Combination Knowledge from Biomedical Literature. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1-7. 2019.
8. Rashmie Abeysinghe, Fengbo Zheng, Eugene W. Hinderer, Hunter N.B. Moseley, Licong Cui. A Lexical Approach to Identifying Subtype Inconsistencies in Biomedical Terminologies. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1982-1989, 2018.
9. Fengbo Zheng, Licong Cui. Exploring Deep Learning-based Approaches for Predicting Concept Names in SNOMED CT. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 808-813, 2018.